Whole Genome Linkage Analysis in a Large Multigenerational Family From Brazil and Case Control Exploration of Linkage Regions

Alsabban, Shaza

Awarding institution:
King's College London

Title: Whole Genome Linkage Analysis in a Large Multigenerational Family From Brazil and Case Control Exploration of Linkage Regions

Author: Shaza Alsabban

Whole Genome Linkage Analysis in a Large Multigenerational Family From Brazil and Case Control Exploration of Linkage Regions

Shaza Issam Alsabban

Submitted for the degree of Doctor of Philosophy in

Social Genetic and Developmental Psychiatry

King's College London, University of London

January 2012

Abstract

Substantial evidence from family and twin studies confirms the importance of genes in influencing susceptibility to Bipolar Disorder (BPD) and Depression. Genome-wide association studies have uncovered a few genetic variants of small effect that explain only a fraction of the total heritability of these disorders, and linkage studies have not been able to identify consistent and replicable findings, possibly due to phenotypic complexity and genetic heterogeneity. Large multigenerational families work as powerful samples to mapping loci for complex diseases as they segregate fewer disease causing genes than a collection of independent nuclear families. These fewer genes segregating may also be more highly penetrant and easier to detect in linkage studies.

This study performed a whole genome linkage scan of a large multigenerational family from Brazil segregating a severe form of BPD and unipolar depression with the aim of localising and identifying genetic variants that contribute to the development of BPD. The 'Brazilian Bipolar Family' (BBF) is one of the largest reported in the literature. Three hundred and eight family members were interviewed using the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID) and the Kiddie-SADS-Present and Lifetime Version (K-SADS-PL) and three-hundred and twenty-four family members were genotyped using the Affymetrix 10K array. Parametric and non-parametric linkage analyses were performed using four hierarchical phenotype models. Four genome-wide significant linkage regions were identified on chromosomes 2p23.1-p22.3, 3p24.3-p24.1, 11p15.4, and 12q24.22-q24.32, and four suggestive linkage regions were identified on chromosomes 1p22.2-p21.3, 1q21.1-q21.3, 12p13.32-p13.31, and 22q11.21-q12.1, which either conferred

specific risk to BPD, unipolar depression, or provided evidence for a general mood disorder liability.

To determine the role of the identified linkage regions in sporadic bipolar and depression cases, I performed a case control association analysis using bipolar and depression case control cohorts. None of the linkage regions identified in the BBF were found to be associated with BPD or depression. The future aim of this project is to determine the functional variants within the identified linkage regions that may be contributing to the development of mood disorders in the BBF through sequencing analysis, which is already underway.

Acknowledgments

I am enormously grateful to my supervisors Professor Peter McGuffin and Dr Gerome Breen for their encouragement, expert guidance and continued support throughout my PhD. I would like to express my appreciation to Dr. Mateus Diniz, Dr. Andiara Rodrigez and Dr. Ary Gadelha who collected the Brazilian Bipolar Family and Professor Rodrigo Bressan for giving me the opportunity to undertake this research and for their continued support throughout. I would also like to thank Dr. Jose Paya Cano for training me in the Affymetrix microarray genotyping system. I would like to thank members of the Depression Consortium Team at the Institute of Psychiatry particularly Professor Anne Farmer and Professor Cathryn Lewis for their continued help and support.

I do not think I would haven been able to complete this thesis without the help and support of my friends at the SGDP, particularly Dr. Margarita Rivera, Dr. Ursula Parades, Dr. Katherine Tansy, Dr. Sarah-Cohen Woods, and Dr. Initi Pedrose your friendships and advice have been invaluable.

I am eternally grateful to my father for constantly encouraging me and supporting me every step of the way and to my brothers for all the fun they bring into my life. Thank you.

Statement of Work

The Brazilian Bipolar Family Study is part of an ongoing collaborative research between the Universidade Federal de São Paulo (UNIFESP) and the Social, Genetic, Developmental Psychiatry research centre. The clinical recruitment and collection of DNA samples was conducted by Dr. Mateus Diniz, Dr. Andiara Rodrigez and Dr. Ary Gadelha under the supervision of Professor Rodrigo Bressan. DNA extraction was performed by Dr. Camila Guindalini and members of her laboratory in UNIFESP. On commencing my PhD, phenotypic and DNA data were already collected. I was involved with scoring the diagnostic interviews with Dr. Mateus Diniz, Dr. Andiara Rodrigez and Dr. Ary Gadelha. The majority of the Affymetrix 10K microarray genotyping was conducted by myself after initial help and training from Dr. Jose Paya Cano. I was responsible for creating and cleaning of the BBF dataset. Genotyping using the 610 QuadBeadChip Illumina array was carried out commercially by deCODE genotyping services.

Data from the Bipolar Case Control (BACC), the Depression Case-Control (DeCC), the Depression Network (DeNT) and the Genome-Based Therapeutic Drugs for Depression (GENDEP) were made available to the author of this thesis by members of the Depression Consortium at the Institute of Psychiatry. In all other respects to the best of my knowledge, the work presented in this thesis is original and my own work.

Content

List of Tables

List of Figures

Chapter 1 General Introduction

1.1 Bipolar Disorder Definition

Bipolar disorder (BPD) is often now considered to be a spectrum of mood disorders that characteristically involve recurrent mania, or milder forms of 'high' mood, and depressive states (Smith et al., 2011). Disorders within this spectrum display a gradation in symptom severity and impairment. The 'classic' varieties of BPD are distinguished in the official nomenclature of Diagnostic and Statistical Manual of Mental Disorders Fourth Edition (DSM-IV) (American Psychiatric Association, 1994) as Bipolar I Disorder (BPI) and Bipolar II Disorder (BPII), characterised by at least one episode of mania and one episode of hypomania respectively. Mania and its less severe form hypomania feature elevated or irritable mood associated with a number of persistent symptoms; inflated self-esteem or grandiosity, distractibility, flight of ideas or racing thoughts, pressured speech, increase in goal directed behaviour, and decreased need for sleep. The two forms of mania are almost identically defined except mania causes impairment in social, occupational, and personal functioning usually requiring hospitalisation, while hypomania does not. Psychotic symptoms, in the form of hallucinations and delusions also differentiate between the two, occurring in mania but never in hypomania. Although some patients with BPD experience only 'highs' the majority also have major depressive episodes, and oscillations between mania or hypomania and depression occur with relatively euthymic periods in between.

Also within the bipolar spectrum are cyclothymia, a persistent mood instability, characterised by episodes of hypomania and mild depression lasting for at least two years with little interruption, and Bipolar Disorder Not Otherwise Specified

(BPNOS), an interesting and nebulous DSM diagnostic category that encompasses bipolar symptoms that do not fit into any of the three bipolar DSM categories described above. BPNOS includes subthreshold manic and hypomanic episodes that do not meet the duration requirement to qualify as actual episodes; manic or mixed episodes superimposed on delusional disorder, residual schizophrenia, or psychotic disorder not otherwise specified; manic or hypomanic episodes without depressive episodes; hypomanic episodes occurring along with chronic depression that are too infrequent to qualify as cyclothymia, or; drug induced manic or hypomanic episodes.

The spectrum of BPD is extended to include schizoaffective disorder, bipolar type, referred to herein and after as SAD. Schizoaffective disorder is a perplexing mental illness that has both features of schizophrenia, including hallucinations, delusions, and distorted thinking, and features of either depression or mania. For a diagnosis of schizoaffective disorder to be given psychosis must be experienced for a period of two weeks without signs of mania or depression. Gershon et al. (1982) in line with Tsuang et al. (1979) advocated the extension of the BPD spectrum to include schizoaffective disorders based on observations from family data where over-transmission of BPD, schizoaffective disorder and unipolar depression was observed in families of probands with mood disorders. Schizoaffective disorder has an interesting history in the DSM where it was initially considered as a subtype of schizophrenia (DSM-I and DSM-II), then as a "Psychotic Disorder Not Otherwise Specified," in DSM-III and as its own separate diagnostic criteria in the revised third edition of the DSM in which the depression and mania subtypes were delineated. In the current edition of the DSM it is categorised under "Schizophrenia and other Psychotic Disorders." In genetic studies, schizoaffective disorder, bipolar subtype (SAD) is often considered a severe subtype of BPD.

1.2 Epidemiology of Bipolar Disorder

It is estimated that broadly defined bipolar spectrum disorders including cyclothymia and BPNOS occur with a lifetime prevalence of 5 to 8% (Judd & Akiskal, 2003) with some studies citing rates as high as 11% (Angst et al., 2003). This estimate is reduced to 1.5% to 3.9% when a narrower definition of bipolarity including only BPI and BPII disorder is considered and to approximately 1% when BPI disorder only is considered (Grant et al., 2005; Kessler, Chiu, Demler, & Walters, 2005; Weissman et al., 1996). These prevalence estimates mirror the increase in symptom severity from BPI disorder to subthreshold bipolar symptoms under the diagnosis of BPNOS, with the more severe BPI disorder occurring with less frequency in the population. There is no evidence of sex differences in the rates of BPI disorder, however, there is evidence that women are over-represented in BPII disorder (Arnold, 2003).

Onset of BPD usually occurs in adolescence or early adulthood, with a mean age of onset around 21 years old (Weissman et al., 1996). Subtypes of BPD have been associated with different ages of onset. A U.S population based study reported an age of onset of 18.2 years for BPI disorder, 20.3 years for BPII disorder and 22.2 years for subthreshold hypomania falling under the umbrella of BPNOS (Merikangas et al., 2007). Further, it is suggested that there are three peak ages of onset for BPD, occurring in adolescence between 14 and 18 years old, in early adulthood between 25 and 27 years old, and in adulthood between 40 and 46 years old (Bellivier et al., 2003; Leboyer, Henry, Paillere-Martinot, & Bellivier, 2005).

The recurrent nature of manic and depressive symptoms in BPD makes it a very debilitating disease. The World Health Organisation reports BPD as the seventh cause of years lost due to disability worldwide, outranking all cancers and primary neurologic disorders, such as epilepsy and Alzheimer's disease, primarily because of

its early onset and chronicity across the lifespan (Mathers, Fat, & Boerma, 2008). BPD is also a leading cause of premature mortality due to suicide with reports of 40 to 45% of sufferers attempting suicide and 15 to 20% completing suicide (Osby, Brandt, Correia, Ekbom, & Sparen, 2001).

1.3 Heritability of Bipolar Disorder

1.3.1 Family Studies

Family studies address whether a disorder clusters in families by comparing the prevalence of the disorder among first-degree relatives of affected probands (cases) to the prevalence in the general population or among relatives of unaffected probands (controls). They also attempt to clarify the genetic overlap between clinical disorders or subtypes of a disorder by following their co-segregation in families (Sham, 1996). It is important to keep in mind that while family studies support the possibility that genes influence a disorder, they cannot establish the role of genes or estimate the magnitude of their influence, as familiality may be caused by environmental, genetic, and gene-environment interactions.

Family studies consistently point to the role of genetics in the aetiology of BPD. Approximately 20 family studies confirm that first-degree relatives of probands with BPD have a 5 to 10 fold increase in risk of developing BPD compared to relatives of controls. The increased heritability is not however limited to BPD, as a three fold increase in the risk of developing unipolar depression is also reported in first degree relatives of probands with BPD compared to controls indicating that it does not “breed true” (Jones, Kent, & Craddock, 2002).

The overlap between bipolar spectrum disorders has been partially clarified by family studies. The main question being, are BPD subtypes distinct entities or are they all variants of the same disorder with similar familial patterns of segregation? Some studies have reported the risk of BPI disorder to be elevated in the families of BPII probands, the risk of BPII disorder to be elevated in the families of BPI probands, and the risk of unipolar depression to be similar in families of BPI and BPII probands (Smoller & Finn, 2003), thus suggesting that the two bipolar subtypes are similar, while other studies have reported that the risk of BPII disorder to be high among families of BPII probands only (Heun & Maier, 1993), suggesting that it is a distinct disorder. The risk of BPI disorder in the relatives of schizoaffective probands has also been reported to be elevated (Andreasen et al., 1987) and vice versa where the risk of schizoaffective disorder has been reported to be elevated in the relatives of BPD (and schizophrenia) probands (Kendler et al., 1993a) attesting to the relatedness of the disorders.

The co-segregation of cyclothymia with the 'classic' subtypes of BPD is documented by a few studies that report the clustering of BPI, BPII, and cyclothymia in families (Akiskal & Akiskal, 1992; Klein, Clark, Dansky, & Margolis, 1988). However, no published studies address the familial overlap between subthreshold manifestations of bipolar disorder and BPI or BPII disorder.

1.3.2 Twin Studies

Twin studies compare the concordance rates of a disorder between monozygotic (MZ) and dizygotic (DZ) twins in an attempt to untangle the genetic determinants of disease from other familial factors. Under the assumption that shared environmental influences on MZ twins are not different from shared environmental influences on DZ twins, higher concordance rates in MZ twins are taken to reflect genetic influences.

Additive genetic influences, shared familial environment (e.g. social class, parents' rearing style), and specific environment (e.g. stressful life events) are used to estimate heritability, defined as the proportion of phenotypic variance attributed to genetic influences, using model fitting (Sham, 1996).

Twin studies have repeatedly documented that the concordance rates for BPD are significantly greater among MZ twins than among DZ twins. The three largest recent twin studies have reported a concordance rate of 38.5% to 43% for MZ twins compared with 4.5% to 5.6% for DZ twins (Kendler, Pedersen, Neale, & Mathé, 1995; Kieseppa, Partonen, Haukka, Kaprio, & Jouko, 2004; McGuffin et al., 2003). Further, estimates of BPD heritability from these studies are high: 93% estimated by Kieseppa et al. (2004) in a population based study of 38 monozygotic twins in Finland, 85% estimated by McGuffin et al. (2003) in a hospital study of 67 twin pairs in the UK, and 83% estimated by Kendler et al. (1995) in a population based study of 35 twins in Sweden. The three studies attributed the remaining variance in liability to BPD, an estimated 7%, 15%, and 17% respectively, to specific non-familial environmental factors. These findings suggest that most of the familiality of BPD could be accounted for by additive genetic effects with no contribution from shared family environment.

1.3.3 Adoption Studies

Adoption studies compare the rates of a disorder in biological family members to those in adoptive family members to distinguish genetic from environmental influences on a disease. If biological family members resemble each other more than adoptive family members then the disorder is influenced by genetics, and if adoptive family members resemble each other more than biological family members then the disorder is influences by environmental factors. Two adoption studies found the

biological parents of adopted-away children with BPD had a higher frequency of affective illness (BPD, unipolar depression, schizoaffective disorder, and cyclothymia) than adoptive parents (Mendlewicz & Rainer, 1977; Wender et al., 1986) and one study showed no evidence of a genetic effect (von Knorring, Cloninger, Bohman, & Sigvardsson, 1983). It is important to note, however, that BPD adoption studies are based on small samples and are therefore regarded as inconclusive.

1.4 The Definition of Bipolar Disorder in Genetic Research

The definition of BPD in genetic research is a debated topic. Some authorities, e.g. Baldessarini (2000) and Soares & Gershon (2000) advocate using narrowly defined BPD, mainly BPI disorder and/or SAD in genetic research as perhaps being more familial than more broadly defined mood disorders. Under this view, sub-dividing BPD into clinically homogenous groups that are reliably assessed either through clinical history or an assessment instrument is performed with the aim of defining a 'tighter' phenotype for genetic studies (McMahon & Schulze, 2005). The familial aggregation of early onset BPD (particularly pre-pubertal-onset) (Geller et al., 2006), episode frequency (Fisfalen et al., 2005), psychosis (Potash et al., 2001), polarity at illness onset (Kassem et al. 2006), rapid cycling (Saunders, Scott, McInnis, & Burmeister, 2008), co-morbid alcohol abuse or dependence (Schulze, Hedeker, Zandi, Rietschel, & McMahon, 2006) are consistent with genetic influences and make them informative sub-phenotypes for BPD studies.

Sub-dividing BPD into more homogenous groups using endophenotypes or phenotypes that are intermediate between the clinical presentation of the disorder and

its genetic underpinnings has received some attention recently. Research based on endophenotypes looks for genes for simple, presumably monogenic, traits that accompany the illness and probably contribute to its pathology. The idea here is to reduce the complexity of disorders such as BPD to more elementary forms (MacQueen, Hajek, & Alda, 2005). Endophenotype that are associated with the illness in the population, have been proven to be heritable, and are manifested regardless of whether or not the illness is active are regarded as suitable. In BPD, promising endophenotypes include abnormal regulation of circadian rhythms, response to sleep deprivation, behavioural response to psychostimulants and other medications, event related potentials, increase in white matter hyperintensities and biochemical alterations in peripheral mononuclear cells have been proposed (Lenox, Gould, & Manji, 2002). However, to date no widely accepted endophenotype for BPD has been discovered.

As opposed to sub-dividing BPD into smaller more homogenous groups, other authorities advocate including more prevalent subtypes of BPD mainly in the form hypomania to facilitate the quest for gene discovery (Akiskal, 2007). This latter view comes with the conviction that all manifestations of BPD are dimensionally rather than categorically different and therefore possess similar aetiological underpinnings.

1.5 Models of Bipolar Disorder Inheritance

A range of models of BPD inheritance have been suggested in the literature, including single major locus models, oligogenic and polygenic models, multifactorial models and mixed models, which will be discussed in this section.

1.5.1 Single Major Locus

Classic Mendelian single locus inheritance assumes that a single genetic defect inherited in a dominant or recessive mode of disease transmission is sufficient to cause disease. Under this model several factors modify the effect of the major locus and render its inheritance pattern more complex; reduced penetrance in which not all individuals carrying the deleterious gene express the disease; pleiotropy where a single genetic mutation could be responsible for variable phenotypic expressions; phenocopies, defined as phenotypes arising from environmental agents that mimic the effects of a mutation in a gene, and; genetic heterogeneity, which refers to situations in which a disease results from a single locus but different loci operate in different families (Ott, 1999). The premise, however, remains that a single locus with a major effect causes disease.

In BPD a single major locus inheritance model has been shown to be mathematically incompatible with available family and twin data (Craddock, Khodel, Van Eerdewegh, & Reich, 1995). To date, no single gene inherited in a dominant, recessive, or intermediate mode has been shown to explain BPD inheritance (Craddock and Forty, 2006). The exception being two segregation analyses that suggest BPD could result from single gene mutations (Rice et al., 1987; Spence et al., 1995). However, segregation analyses tend to be very sensitive to mode of ascertainment, which is almost impossible to allow for correctly (Suarez, Hampe, & Van Eerdewegh, 1994) and are seldom based on large enough datasets to account for the large number of parameters required to model complex diseases (Ott, 1999), so interpretation of these two findings should be conducted with caution.

1.5.2 Polygenic Inheritance

Polygenic inheritance in which multiple genes at different loci contribute to disease susceptibility has been suggested as a likely theory for BPD inheritance (Purcell et al., 2009; Sklar et al., 2011). Under this model, the effects of the genes are additive and mutually independent and the phenotype results from the sum of the risk alleles (McGuffin, 1991). Here one assumes the presence of an underlying quantity or liability to develop the disorder, if it exceeds a certain threshold an individual becomes affected. This model of inheritance, originally postulated by Falconer, (1965) assumes that liability to disease has a normal distribution and the proportions of affected individuals in the population are those whose liability exceeds some threshold. For example, if we assume that multiple BPD causing genes are segregating in the population, and bipolar symptoms are only observable with the cumulative effect of twenty risk alleles, then affected individuals would be those who carry twenty or more risk alleles and unaffected individuals would be those who carry a range of risk alleles, up to nineteen, without developing BPD symptoms. Under a highly polygenic disease where many genetic variants of small effect are involved in disease susceptibility, each affected individual may carry a unique set of risk alleles so that any two affected individuals are highly unlikely to carry the same combination of risk alleles (Visscher, Goddard, Derks, & Wray, 2011).

The variation in symptom severity observed in patients with BPD is compatible with a polygenic model of inheritance with multiple thresholds of liability (Reich, Cloninger, & Guze, 1975), where more severe subtypes of BPD occur at more extreme thresholds, i.e. larger quantities of the same underlying genetic liability. Thus affected individuals with bipolar spectrum disorders would lie beyond the first threshold, e.g. of twenty risk alleles, and individuals affected with more severe forms

of BPD would exceed a more extreme threshold, e.g. of thirty risk alleles for BPII disorder and forty risk alleles for BPI disorder (Figure 1.1)

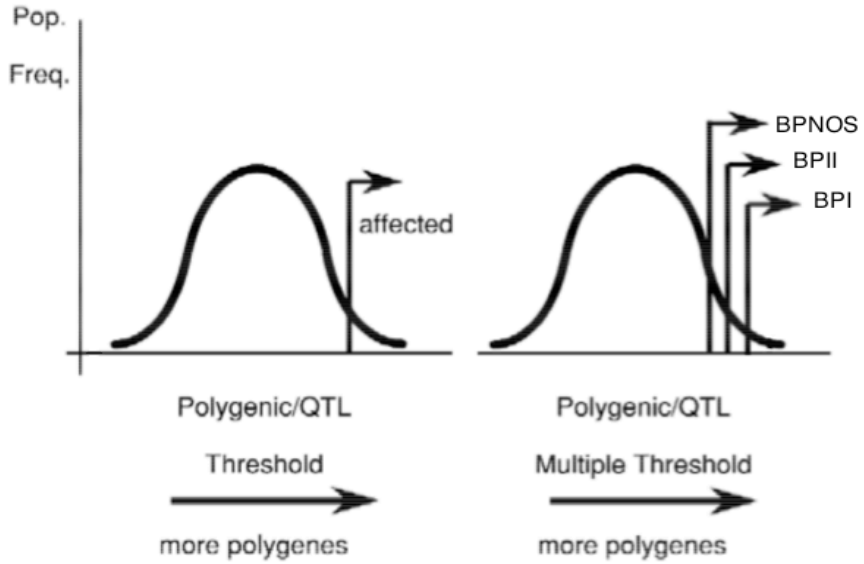


Figure 1.1 The liability threshold and multiple threshold models. The figure on the left depicts a liability threshold model in which a disease results when a critical underlying liability threshold is exceeded. The figure on the right depicts a polygenic multiple threshold model under which different thresholds lead to different subtypes of BPD. This models holds that the greater the number of risk alleles one has the greater the severity of the disease. A gradation in the underlying disease liability similar to that observed in the clinical manifestations of the disease is from subthreshold mania and hypomanic, to BPII disorder to BPI disorder. Graph adapted from Kelsoe et al. (2003).

1.5.3 Multifactorial Inheritance

The multifactorial inheritance model proposes that genes and environmental factors combine to cause disease. The environmental factors in question could be due to familial environment (e.g. low socioeconomic status), specific environment (e.g. death of a spouse), gene-environment correlations and/or gene-environment interactions. The role of the latter two in susceptibility to psychiatric disorders has received increasing attention in recent years. In gene-environment correlations, genes act by influencing the probability that an individual will be exposed to a life event. For example, McGuffin, Katz, & Bebbington (1988) found not only an increased rate of depression among relatives of depressed probands but also an increased reporting

of life events. This finding was replicated by several studies that showed life events aggregated in families (Kendler, Neale, Kessler, Heath, & Eaves, 1993b; Rijsdijk et al., 2001). Other studies, however, did not find significant difference between the number of threatening life events experienced by the siblings of individuals with depression and the siblings of healthy controls (Farmer et al., 2000).

In gene-environment interactions, also referred to as 'genetic control of sensitivity to the environment,' coined by Kendler (1998), genes modify the pathogenic effects of the environment to make an individual more or less vulnerable to disease. So far gene-environment interactions have only been tested on the level of candidate genes. In a landmark study, Caspi et al., (2003) explored gene-environment interactions in the Dunedin birth cohort and found a functional polymorphism in the promoter region of the serotonin transporter gene *(5-HTT)* moderated the influence of stressful life events on depression. Similarly, a valine to methionine substitution polymorphism in the *BDNF* gene was found to moderate the effects of stressful life events on the propensity to develop depressive (but not manic) episodes in BPD patients (Hosang et al., 2010). The role of gene-environment correlations and interactions in BPD, particularly mania is not well studied. However, based on empirical findings the inheritance of BPD could be expressed as follows:

BPD phenotype = genotype + environment + gene-environment correlation + gene-environment interaction

1.5.4 Mixed Model of Inheritance

The mixed model of inheritance, which refers to the effect of one or more major loci and oligogenes (few genes) or polygenes on disease susceptibility (Faraone, Tsuang, & Tsuang, 1999) has received some support in BPD. Cystic fibrosis is a good

example of a disease driven by mixed inheritance. A mutation in the *CFTR* gene causes cystic fibrosis, however, phenotypic variations observed in patients depend on a discrete number of alleles at different loci (Badano & Katsanis, 2002). BPD is proposed by Ewald, Kruse, & Mors (2003) and Ewald et al. (2005) to occur due to dominant risk at a major locus operating with a number of risk and modifying genes inherited in a recessive mode at other loci. This suggestion is based on the authors' identification of a number of recessively inherited loci that segregate with BPD in pedigrees from Denmark and Cuba together with findings from the literature that implicate dominantly inherited loci in BPD. The presence of a major locus that is co-inherited or modified by recessively inherited risk was not however examined, so the suggestion remains speculative.

1.5.5 Epistasis

The Epistasis model suggests that a system of interacting genes may be liable in BPD with interactions between genes of small effect or between genes of small effect and a major locus (Craddock & Forty, 2006). Consequently, standard statistical methods that treat disease loci as if they were independent of each other may not have sufficient power to detect genes of modest effect. The simultaneous consideration of multiple genomic loci in linkage analysis has yielded convincing results of the applicability of epistatic models to complex disorders in general, for example in type 1 diabetes Cox et al., (1999) employed a locus interaction model to demonstrate that accounting for locus interactions increased evidence for linkage (to chromosome 15) from marginal to significant, and in BPD Fullerton, Donald, Mitchell, & Schofield (2010) identified three significant linkage peaks and four gene interaction clusters in a study of 65 Australian extended BPD pedigrees, thus emphasising the role of genetic interactions in BPD aetiology.

1.6 Molecular Genetic Investigation of Bipolar Disorder

The evidence for a genetic contribution to BPD from family, twin and adoption studies spurred numerous efforts to localise and identify the genetic variants that contribute to the development of BPD using whole genome linkage and association studies, and more recently whole genome sequencing. These methods although methodologically different offer complementary tools for understanding the genetic architecture of complex diseases, and are all necessary for gaining insight into the pathophysiology of diseases with implications for genetically informed therapeutics and disease prevention.

1.6.1 Linkage Studies

Linkage is very often the first stage in the molecular genetic investigation of a disorder. It is used to identify genomic region(s) that may contain disease genes (or loci) by following the co-segregation in informative pedigree(s) of the disease of interest and genetic markers such as microsatellites or single nucleotide polymorphisms (SNPs) with known locations in the genome (Sham & McGuffin, 2002). A large number of nuclear pedigrees or a small number of multigenerational extended pedigrees with multiply affected members are studied with the aim of mapping genetic variants with comparatively large effect sizes typically extending between 5 to 20 centiMorgans (cM). By definition linkage is the tendency of genes to be transmitted together from parent to offspring more often than expected, thus representing a departure from Mendel's law of independent assortment. Linkage is the main topic of this thesis and will be discussed thoroughly in chapter 3.

1.6.1.2 Linkage Findings in Bipolar Disorder

Over 40 whole genome-linkage studies have been published for BPD with no major and consistent 'hits' found and multiple regions of the genome implicated in the disorder. Three meta-analyses have been performed in an effort to make sense of the large, and potentially perplexing data. The first by Badner & Gershon (2002) combined data from seven whole genome scans for BPD. The authors used a method derived from Fisher's approach to combining p-values with a correction for the size of the linkage regions included. They found two significant susceptibility loci for BPD on chromosomes 13q and 22q. A second and more extensive meta-analysis studied linkage statistics from eighteen BPD whole-genome scans and found no genome wide significant regions and only reported suggestive linkage on chromosomes 9p, 10q, 14q, and 18 (Segurado et al., 2003). The most recent major combined analysis by McQueen et al. (2005) was a 'mega-analysis' since it used the original genotype data from eleven BPD genome studies including 5,179 individuals from 1,067 families. The study found significant linkage to chromosome 6q for BPI disorder and chromosome 8q for BPI and BPII disorders. In addition, several consortia have combined data across multiple sites in an attempt to identify "truly" positive linkage regions. Lambert et al., (2005) combined data from the UK and Ireland and found support for regions on chromosome 9p21, previously identified in the meta-analysis by Segurado et al. (2003) and 10p14-21. Schumacher et al., (2005) found evidence for linkage on chromosomes 6q24 and 4q31 using combined data from Spanish, Romanian and Bulgarian families and proposed these two previously identified regions should be considered confirmed BPD susceptibility loci.

Many more linkage regions have been implicated in BPD. Linkage peaks identified with genome wide significance (defined by LOD scores greater or equal to

3.0) in independent studies are presented in Table 1.1. Only regions supported by subsequent linkage hits are presented.

Region	Publication	Phenotype	Model	Sample	Maximum LOD
4p16	Blackwood et al. (1996)	BPI, BPII	Dominant	Large Scottish Pedigree	MLOD=4.8
5q33-q34	Herzberg et al. (2006)	BPI	NPL	Costa Rica, Colombia	LOD=4.18
	Jasinska et al. (2009)	BPI	NPL	Costa Rica, Colombia	LOD=4.90
6q16.3	Dick et al. (2003)	BPI, BPII, SAD, or UP	ASP	NIMH-II	LOD=3.61
6q22	Middleton et al. (2004)	BPI, SAD	NPL	Portuguese	LOD=3.24*
6q23-q24	Venkon et al. (2005)	BPI, BPII, SAD, UP	Recessive	Västerbotten	MLOD=3.25
8p21	Park et al. (2004)	BPI, SAD	Dominant	Large Pedigree	LOD=3.46
9q31	Park et al. (2004)	BPI, BPII, SAD,UP,SAD	Recessive	Large Pedigree	LOD=3.55
9q31-q33	Venkon et al. (2005)	BPI, BPII, SAD, UP	Recessive	Västerbotten	MLOD=3.70
10p12	Rice et al. (1998)	BPI, BPII, SAD	ASP	NIMH-I	MLOD=3.40
10q22.2	Rice et al. (1998)	BPI, BPII, SAD	ASP	NIMH-I	MLOD=3.47
11p11	Middleton et al. (2004)	BPI, SAD	NPL	Portuguese	LOD=4.27*
12q21.33-q24.23	Morissette et al. (1999)	BPI, BPII, SAD, or UP	NPL	Québec	LOD=3.63
12q24.32	Ewald et al. (2002)	BPI, BPII, SAD	NPL	Two Danish families	LOD=3.42
13q31	Detra-Wadleigh et al. (1999)	BPI, BPII, SAD, or UP	ASP	U.S pedigrees	LOD=3.50
16p13.11	Mérette et al. (2008)	BPI, BPII, UP	NPL	Eastern Quebec	LOD=3.87*
16p13.11	Ross et al. (2008)	BPI, BPII	Dominant	NIMH-all	LOD=3.20
17q25.3	Dick et al. (2003)	BPI, BPII, SAD	ASP	NIMH-II	LOD=3.63
18q21-23	Stine et al.(1995)	BPI, BPII, SAD, or UP	Dominant	U.S pedigrees	LOD=3.51
21q22	Straub et al. (1994)	BPI, BPII, SAD, or UP	Dominant	Large U.S Pedigree	LOD=3.41
22q11-12	Potash et al. (2003)	Psychosis	NPL	Baltimore, Iowa	LOD=3.06
	Kelsoe et al. (2001)	BPI, BPII, SAD	Dominant	North America	LOD=3.84

Table 1.1. Significant Linkage peaks identified in independent BPD studies. BPI=bipolar I disorder, BPII=bipolar II disorder, SAD=schizoaffective disorder, bipolar type, UP=Unipolar Depression, NPL=non-parametric linkage, MLOD=Multipoint LOD score. An asterisk denotes NPL scores that had been converted to LOD scores using LOD=NPL^2/4.62 (Ott, 1999). Although these regions are not inclusive of all the significant linkage findings reported in the literature, they are linkage peaks that have been confirmed by subsequent linkage studies and are presented to demonstrate the variability in linkage findings reported in the literature.

1.6.2 Association Studies

Genetic association studies are used to evaluate the relationship between genetic variants and the risk of developing disease by comparing individuals with a disease (cases) with unaffected individuals (controls) from the same population for differences in genotype or allele frequencies. They are based on linkage disequilibrium (LD), which refers to the non-random association between alleles at adjacent genetic loci such that certain combinations of alleles are more likely to occur together on a chromosome than other combinations of alleles. If we consider two biallelic markers, AG and CT, there are four possible combinations of alleles or haplotypes, AC, AT, GC, and GT. Under linkage equilibrium, the alleles in a haplotype are randomly associated with each other and the population frequency of each haplotype is equal to the frequencies of the alleles it contains. For example, the population frequency, P ,of haplotype AC would be simply defined as follows

$$P_{AC} = P_A P_C$$

Where P_{A} is the population frequency of allele A at marker AG and P_C is the population frequency of allele C at marker CT. Linkage Disequilibrium refers to deviations from the expected haplotype frequencies in the population due to the non-random association between alleles at adjacent loci. As such when an allele at one locus is found to occur more often than expected by chance with another allele at another locus, the frequency of that haplotype is increased (Ardlie, Kruglyak, & Seielstad, 2002) (Figure 1.2).

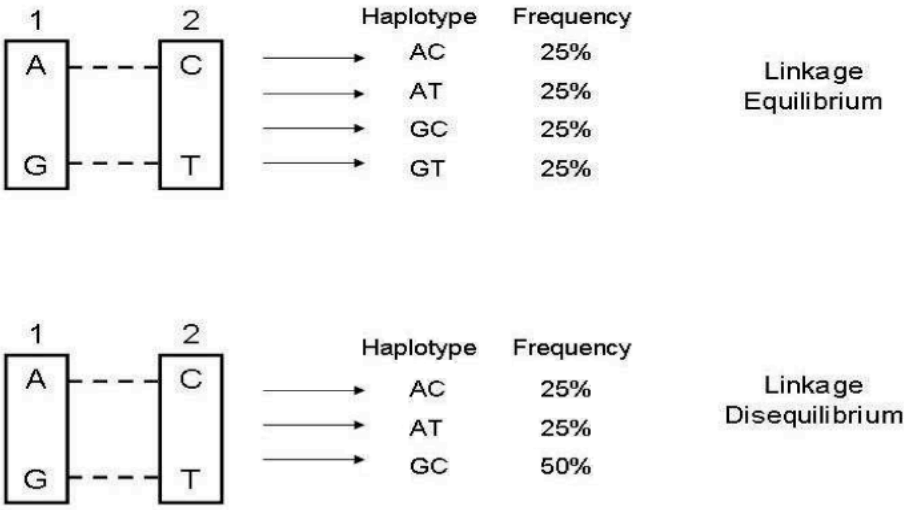


Figure 1.2 Linkage equilibrium and linkage disequilibrium in two biallelic markers. Linkage equilibrium occurs when all the observed haplotypes follow the expected population frequencies, which in this example are 25% for each haplotype. Linkage Disequilibrium occurs when deviations from the expected frequencies are observed in the population.

An important distinction between linkage and LD needs to be made. Two genetic loci are linked if they are transmitted from parent to offspring more often than expected under independent assortment. They are in LD if across the population as a whole they are found together on the same haplotype more often than expected. An association between a specific polymorphism (e.g. SNP) and disease susceptibility allele(s) indicates that the two are so close together that they are rarely separated by meiosis. Two loci in LD will also be in linkage but not vice versa. Because linkage extends over large regions of the genome it detects genes of major effect, LD on the other had occurs over relatively small genetic distances, generally not extending over regions of one centiMorgan in outbred populations and is therefore used to map susceptibility genes with small effect sizes (Reich et al., 2001)

Until recently association studies have been limited to identifying candidate genes thought to influence disease susceptibility because of what is known about their neurochemistry and/or mechanisms of drug action and positional candidate genes

implicated because they lie within significant linkage peaks. Serretti & Mandelli, (2008) reviewed published case control association studies conducted until 2007 and reported that candidate genes that have been consistently associated with BPD fall into four different pathway; the serotonin (*5HTT* and *TPH2*), the dopamine (*DRD4* and *DTNBP1*), the glutamate *(DAOA* and *DTNBP1)* and cell growth and/or maintenance pathways (*NRG1*, *DISC1* and *BDNF*). While these and many more candidate and positional genes have been associated with BPD, none have been unambiguously identified as causal.

1.6.2.1 Genome Wide Association Studies

Genome wide association studies (GWAS) at the level of individual genotyping have only become feasible with the advent of microarray technology that allows genotyping of hundreds of thousands of single nucleotide polymorphisms (SNPs) in a single experiment. Based on the "common disease, common variant" hypothesis, GWAS assume that many genes of small effect, but high minor allele frequency, contribute to the liability of common diseases. In fact, most common variants identified from GWAS for a wide range of diseases confer relatively low risk to disease with odds ratios of 1.1 to 1.5 (Bodmer & Bonilla, 2008).

Association methods are compromised by allelic heterogeneity or differences in marker allele frequencies due to population stratification, which refers to the problem of hidden, or cryptic, sub-populations being represented in case and control groups. It can lead to false positive results as allele frequencies may differ between cases and controls irrespective of disease status. As long as cases and controls are well matched for broad ethnic backgrounds and measures are taken to exclude individuals whose genome wide association data reveal substantial differences in genetic background, population stratification can be overcome. Several tools currently

exist to detect and adjust for population stratification, which are primarily based on principal component analysis (Price et al., 2006; Zheng, Freidlin, & Gastwirth, 2006). Assessing population stratification and correcting any differences is part of the primary analysis of most GWAS undertaken.

1.6.2.2 Genome Wide Association Findings in Bipolar Disorder

The first GWAS study of seven common disorders, of which BPD was one, was the Wellcome Trust Case Control Consortium study (2007). To the disappointment of many this only revealed one major "hit" in BPD, which was a SNP on chromosome 16p12 (Burton et al., 2007). This disappointment was carried through when a second large GWAS failed to identify any SNPs with genome wide significance in BPD (Sklar et al., 2008). Subsequent analyses yielded significant evidence for association to a few genes including *CACNA1C,* a gene located on chromosome 12p13 which encodes an L-type, alpha-1C subunit of a voltage-dependent calcium channel protein, which mediates the influx of calcium ions into the cell in response to membrane depolarisation, and *ANK3*, a gene located on chromosome 10q21 involved with cell motility, activation proliferation, contact and the maintenance of specialised membrane domains and is known to modulate the activity of neuronal sodium channels (Ferreira et al., 2008; Smith et al., 2009). More recently, The Psychiatric Genomic Consortium collaboration (PGC) across more than 200 scientists in 65 institutions and 19 countries, and 11,974 cases and 51,792 controls for BPD published a report identifying a new intronic variant in *ODZ4*, located on chromosome 11q14.1, which according to the authors is involved in cell surface signaling and neuronal pathfinding. The study also confirmed the role of *CACNA1C* in BPD (Sklar et al., 2011). These results are a step forward towards unravelling the genetic determinants of BPD. However, to date the identified loci only explain 1 to 2% of the disease

liability (Badner et al. 2011). Further, GWAS have only identified few common risk variants in BPD. Is that a reflection of the limited number of such variants in BPD, or are GWAS not larger enough. The answer seems to be yes and no. On the one hand, the challenge of GWAS remains to be one of sample size. Using large enough samples to perform replication studies to achieve unequivocal significance while accounting for population stratification and multiple testing remains a priority. On the other hand, emerging views hold that complex disorders are actually highly heterogeneous with many uncommon mutations having large effects (Cirulli & Goldstein, 2010) and recent advances in next-generation sequencing technologies should help unravel their genetic architecture.

1.6.3 Copy Number Variants (CNVs)

CNVs are genomic deletions and duplications that range in size from one kilobase (kb) to several megabases (Mb). They may be rare causal mutations or common variants, which increase susceptibility to disease. CNVs have been reported in psychiatric phenotypes such as autism, mental retardation, and schizophrenia (O'Donovan, Craddock, & Owen, 2009). The overall number of CNVs is reported to be greater in individuals with schizophrenia compared with controls, and a number of specific rare CNVs have been associated with schizophrenia, for example deletions at chromosomes 1q21.1, and 15q13.3, 22q11 as well as deletions of the gene encoding the synaptic neural adhesion molecule, neurexin (Kirov et al., 2009). Further, studies show the risk of disease associated with CNVs far exceeds that associated with most common variants identified from GWAS. For example, three rare deletions at 1q21.1, 15q11.2 and 15q13.3 associated with schizophrenia have odds ratios estimated at 2.7, 11.5 and 14.8 (Stefansson et al., 2009), substantially higher than odds ratios for

common SNPs associated with schizophrenia estimated at below 1.3 (Purcell et al., 2009).

The role of CNVs in the aetiology of BPD is yet to be determined. Emerging evidence suggests they influence risk for BPD. A study by Zhang et al. (2008) found singletons deletions greater than 100 kb in length are present in 16.2% of BPD cases in contrast with 12.3% of controls and a meta-analysis of multiple psychiatric disorders showed a 16p11.2 micro-duplication to be associated with BPD. Nevertheless, CNVs appear to contribute less to the susceptibility to BPD than to schizophrenia. Variants influencing BPD seem to be smaller, less likely to be deletions, and have smaller effect sizes (Grozeva et al., 2010; McQuillin et al., 2011).

Currently, the role of CNVs in unaffected individuals is not clearly understood. Larger CNV studies need to be undertaken, particularly in control populations, to determine the average rate and size of CNVs in unaffected individuals. Without knowledge of the affect of CNVs in a control population, conclusions about their role in the aetiology of disease cannot be accurately determined. Molecular studies of CNVs are needed to clearly understand how each individual CNV may give rise to disease.

1.6.4 Sequencing Studies

The recognition that common genetic variants captured using GWAS only explain a minor fraction of the genetic risk of most diseases spurred renewed interest in rare variants and rare CNVs as plausible in disease aetiology and with more reasonable costs, sequencing approaches have recently become the poster for uncovering the missing heritability in complex diseases, particularly rare disease causing variants. Today whole genome sequencing offers a comprehensive collection of rare variants

and structural variation for study and is believed to be the future of molecular genetic investigation of disease.

Given that rare and common genetic variants likely contribute to the genetic architecture of BPD, linkage, association, and sequencing methods are needed to map the full range of BPD susceptibility. Putting already established common variants from GWAS studies together with findings on rare and structural variants from sequencing studies on the verge of delivery to the scientific world will shed more light on the genetics under pinning of BPD and related disorders.

1.7 Differences and Similarities Between Bipolar Disorder and Unipolar Depression

The term unipolar depression refers to three disorders; Major Depressive Disorder Major (MDD), dysthymia, and depression not otherwise specified, which vary in severity and duration. MDD is characterised by persistent negative emotions and thoughts that coexist with disturbances in sleep, energy, and motivational behaviour. It results in perseverative and intrusive thoughts of death, suicide and guilt. For an individual to receive a diagnosis of a major depressive episode on the basis of the DSM-IV at least five of the following symptoms must be present for a minimum period of two weeks accompanied by either depressed mood or loss of interest or pleasure: increased/decreased appetite, insomnia/hypersomnia, psychomotor agitation/retardation, fatigue/loss of energy, feelings of worthlessness and inappropriate guilt, reduced concentration, suicidal behaviour, and anhedonia. Dysthymia, on the other hand, is a chronic, less severe, form of depression that persists for a period of at least two years and depression not otherwise specified is any form of depression that does not fall into the two other forms of depression described.

Kraepelin originally perceived BPD and unipolar depression as a single construct of manic-depressive illness (Angst & Marneros, 2001). The proposal that unipolar depression and BPD should be divided was originally put forward by Leonhard (1959) and seems to be gaining widespread acceptance based on evidence from genetic studies. However, a degree of genetic overlap between BPD and unipolar depression is evident. There is consistent evidence of an increase in the frequency of both BPD and unipolar depression in the close relatives of individuals with BPD and as much as a three-fold increase in the average risk of MDD is reported in relatives of individuals with BPD compared to controls (Jones et al., 2002;

McGuffin & Katz, 1989; Merikangas & Yu, 2002). However, the opposite is not true; relatives of individuals with MDD do not have significantly more BPD than the relatives of controls (Winokur, Coryell, Keller, Endicott, & Leon, 1995). One explanation is proposed by the multiple threshold model, an extension of the liability threshold model *(see section 1.52)* that holds the two disorders exist on the same continuum of liability but are caused by different quantities of the same underlying genetic liability. Under this model all affected individuals lie beyond the first threshold for being affected with unipolar depression, those who exceed a more extreme threshold are diagnosed with BPD, and those who exceed an even more extreme threshold are diagnosed with schizophrenia (Kelsoe et al.,2003) (Figure 1.3).

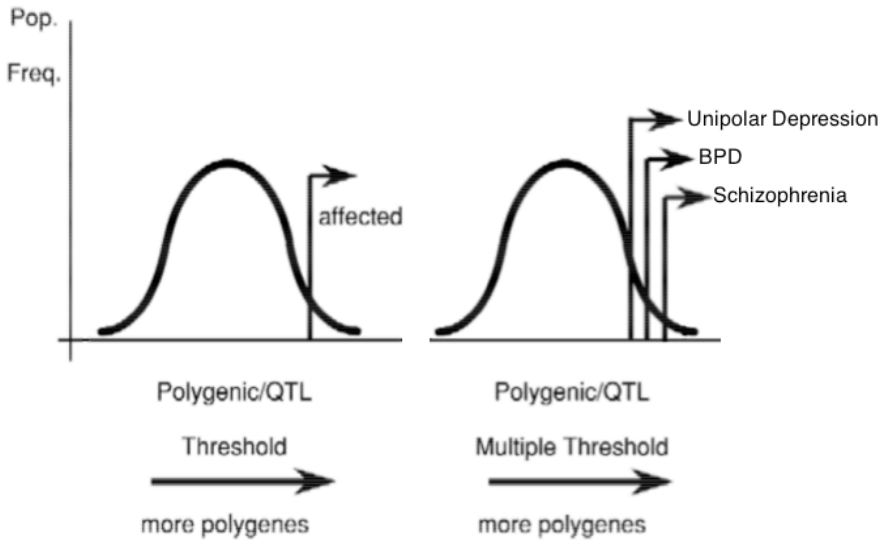


Figure 1.3. The multiple threshold model of mood disorders and schizophrenia. The figure on the left depicts a liability threshold model in which a disease results when a critical underlying liability threshold is exceeded. The figure on the right depicts a polygenic multiple threshold model under which different thresholds lead to different disorders. This models holds that the greater the number of risk alleles one has the greater the severity of the disease. A gradation from unipolar depression, to BPD to schizophrenia is shown. Graph adapted from Kelsoe et al. (2003).

McGuffin et al. (2003) refuted the multiple threshold model in an analysis based on twins ascertained from the Maudsley hospital twin register via probands who received a diagnosis of BPD or MDD. The authors proposed an alternative correlated liability model, which assumes that there are three sources of variance (additive genetic, shared environmental, non-shared environmental) and allows testing for common factors that affect both MDD and BPD and specific factors that may affect one disorder alone. McGuffin et al. (2003) presented evidence that a correlated liability model with both overlapping genetic effects on mania and depression and genetic effects specific to mania provided a good fit. In this study, the overlapping genetic effects on mania and depression were estimated at approximately 29%, indicating that most genetic liability to mania (approximately 71%) was specific to bipolarity (Figure 1.4).

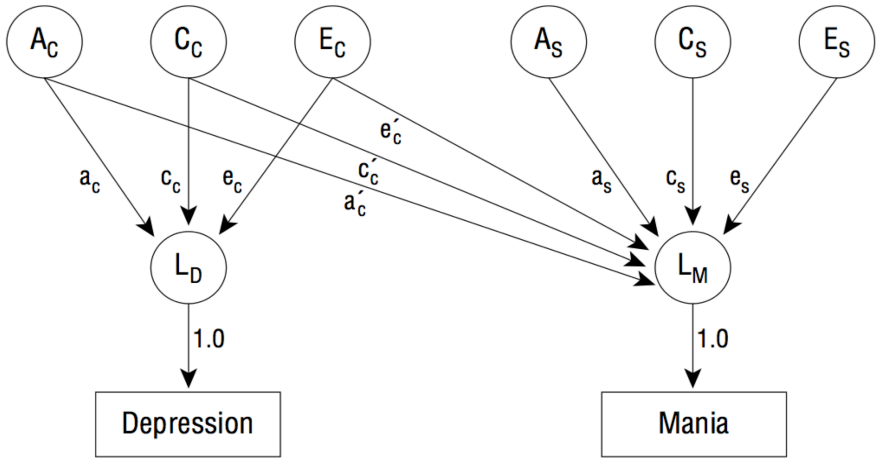


Figure 1.4. Path diagram for a correlated liability model. The correlation between depression and mania is partitioned into additive genetics (A), common environmental (C), and specific environmental (E) in Cholesky decomposition for A, C, and E there are both common factors affecting the disorders and specific factors affecting the second disorder only. The effect of common additive genetic factors A_C on the first disorder is represented by the path coefficient a_C, and the effect on the second disorder is represented by a'_C. The effect of specific additive genetic factors, A_S on the second disorder is represented by the path coefficient A_S. L indicates liability.

Support for a shared genetic liability between BPD and unipolar depression from molecular genetic studies is limited to a few linkage regions and unconfirmed candidate genes. Compared with BPD and schizophrenia a few genome wide scans of unipolar depression as the main phenotype are conducted and there have been no meta-analyses undertaken. In the last years, the first GWAS in MDD were published. None reported genome wide significant results and their top hits were difficult to replicate (Lewis et al., 2010; Muglia et al., 2010; Shyn et al., 2011; Sullivan et al., 2008; Wray et al., 2010).

One region of interest that emerges from linkage studies is on chromosome 12q23-q24, which was implicated in BPD by four genome wide significant signals (Curtis et al., 2003; Maziade et al., 2001; Morissette et al., 1999; Venken et al., 2005) two pedigrees that segregated both BPD and an autosomal dominant skin disorder, Darier's disease (Green et al. 2005), a large sibling pair genome scan of unipolar depression (McGuffin et al. 2005), and a whole genome linkage scan of 1,890 individuals from 110 Utah pedigrees with a strong family history of major depression (Abkevich et al., 2003). A second region is on chromosome 2q33-q36 has been reported in families from a Northern Swedish Isolated population with BPD or unipolar depression (Venken et al., 2005), which overlaps with a linkage signal on 2q33.1 found in 81 families identified by probands with recurrent, early-onset, MDD (Zubenko et al. 2003). Furthermore, candidate gene studies have confirmed some overlap in the candidate genes associated with BPD and unipolar depression. . For example, Brain Derived Neurotrophic factor (*BDNF)* (Groves, 2007), P2X purinoceptor 7 gene (*P2RX7)* (McQuillin et al. 2009), and *GRIK4* (Paddock et al., 2007; Pickard et al., 2006) have been reported in both BPD and depression. However, to date their roles have not been confirmed in either unipolar depression or BPD.

1.8 Differences and Similarities Between Bipolar Disorder and Schizophrenia

BPD and schizophrenia are generally considered distinct nosological entities. However, the two disorders share psychotic features in common and are epidemiologically similar, both show no gender differences, and have similar population prevalences and age of onset distributions. BPD and schizophrenia have been reported to 'breed true' (Frangos, Athanassenas, Tsitourides, Katsanou, & Alexandrakou, 1985; Gershon et al., 1982). However, families in which multiple cases of BPD and schizophrenia co-occur challenge the Kraepelinian dichotomy and premise that families segregate either BPD or schizophrenia. The fact that relatives of probands with BPD and schizophrenia are at an increased risk for schizoaffective disorder and unipolar depression (Kendler et al, 1993b; Maier et al., 1993) and that relatives of probands with schizoaffective disorder are at an increased risk for BPD, unipolar depression and schizophrenia (Rice et al., 1987) question the diagnostic divide between mood disorders and schizophrenia. A twin study of 77 MZ and 89 same sex DZ twins recruited through probands with schizoaffective symptoms indicated, using an analysis unconstrained by traditional diagnostic hierarchies (that is, the principal that schizophrenia "trumps" mood disorder in diagnosis), that there are common genes that confer susceptibility to both BPD and schizophrenia as well as genes that are specific to each of these disorders. The study also found that the genetic liability to schizoaffective disorder was entirely shared in common with schizophrenia and BPD (Cardno, Rijsdijk, Sham, Murray, & McGuffin, 2002). This study was based on the correlated liability model described in *section 1.7*

Molecular genetic studies have shown some genes common to both BPD and schizophrenia supporting a BPD-schizophrenia overlap. Chromosomes 1q42, 10p11-

16, 13q32, 18p11 and 22q11-13 have been linked to both disorders with overlapping or convergent evidence (Baron, 2002). In addition, a genome wide linkage scan using families with BPD or schizophrenia, ascertained through probands with schizoaffective disorder, demonstrated significant linkage at 1q42 and suggestive linkages at 22q11 with equal contribution to the LOD score from 'schizophrenia' and 'bipolar' families (Hamshere et al., 2005). Further, a number of candidate genes have been shown to influence susceptibility to BPD and schizophrenia. The 'disrupted in schizophrenia', *DISC1*, locus at 1q42 is implicated in schizophrenia, BPD, and schizoaffective disorder by linkage studies and evidence for allelic association at polymorphisms at this locus are reported in schizophrenia, BPD, and schizoaffective disorder (Craddock, O'donovan, & Owen, 2005). Neuregulin 1 *(NRG1),* one of the first confirmed genes conferring susceptibility to schizophrenia, initially identified in 33 Icelandic families and later replicated in Scottish, Welsh, and Chinese families has been shown to have a similar effect with BPD (Farmer, Elkin, & McGuffin, 2007). Genetic variation at the *COMT* gene, which lies on locus 22q11, is shown to confer susceptibility across the psychosis spectrum in both BPD and schizophrenia. More recently, the PGC, in a combined GWAS analysis of schizophrenia and BPD found strong association evidence for SNPs in *CACNA1C* and in the region of NEK4-ITIH1-ITIH3-ITIH4 (Sklar et al., 2011), indicating their involvement in both disorders.

Chapter 2 Subjects

This chapter outlines the structure and phenotypes found in a single large multigenerational family from Brazil used for the whole linkage analysis conducted in this thesis.

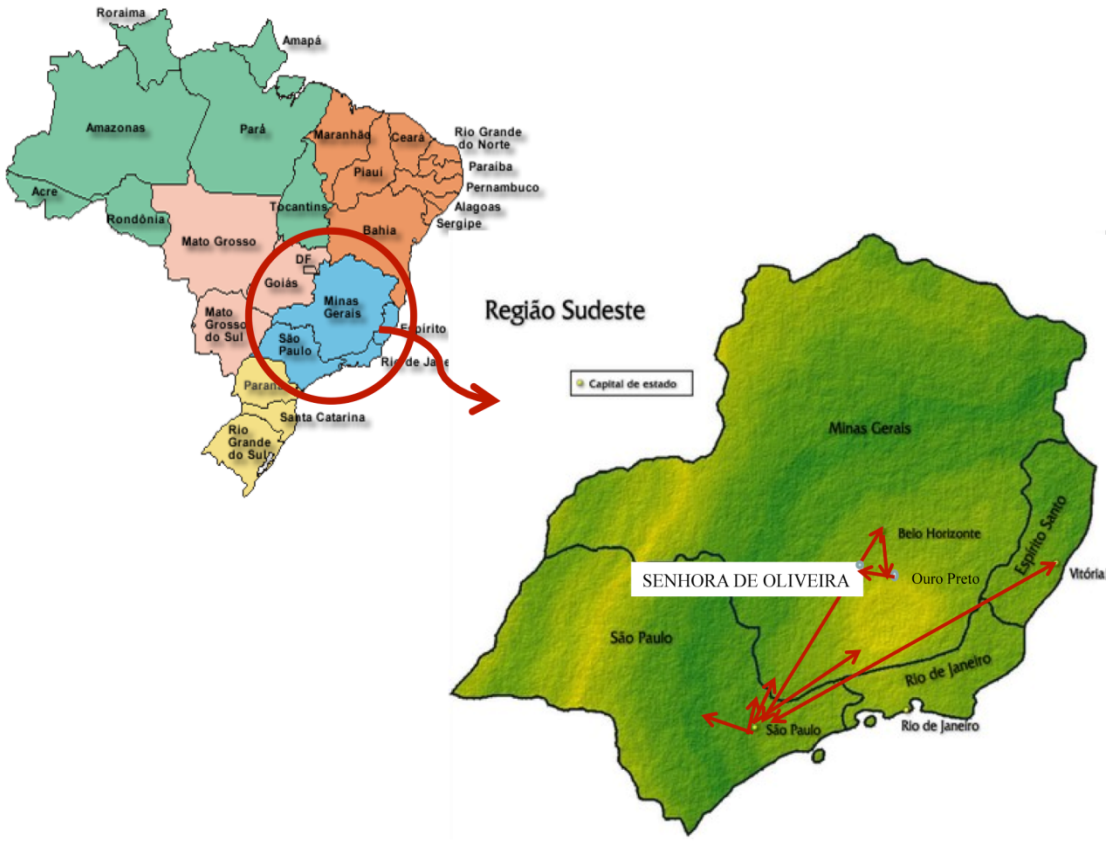


Figure 2.1. Map of Brazil (left top) highlighting South East Brazil, shown in greater detail in the bottom right. Most family members were recruited from Senhora De Oliveira, a small municipality in Minas Gerais, or São Paulo and Vitória where some family members moved. The red arrows point to the locations visited by the research team to recruit the family.

2.1 Subjects for the Whole Genome Linkage Study: The Brazilian Bipolar Family (BBF)

The Brazilian Bipolar Family (BBF) is a five-generation family of 639 members, 57 of whom are confirmed to be deceased. The family originates from Senhora de Oliveira, a small Brazilian municipality located in the state of Minas Gerais in South Eastern Brazil. Senhora De Oliveira has a population of approximately 6000, according to a 2004 estimate, and was founded in 1695 (Wikipedia).

The area is relatively underdeveloped with no paved roads leading to the town and little in the way of formal healthcare available. Three hundred and thirty family members (49% males and 51% females) were recruited by three psychiatrists from the University of São Paulo between 2006 and 2008 from Senhora de Oliveira and surrounding cities in the state of Minas Gerais, São Paulo in the state of São Paulo, and Vitória in the state of Espírito Santo. The majority of the family still resided in Senhora de Oliveira or nearby cities, where they are widely known to be "eccentric and crazy." A subset of the family, mainly young professionals from the third generation moved with their nuclear families to either São Paulo or Vitória (Figure 2.1).

2.1.1 Family Ascertainment

The BBF was ascertained through a 45-year-old female proband with severe BPI disorder who was treated by one of the psychiatrists involved in the study. The grandparents of the proband (201, 200) are reported to be first cousins, and although deceased, the proband's grandmother (200) has hospital records indicating she suffered from BPI disorder, and the proband's grandfather (201) is suspected of being on the bipolar spectrum. The children, grandchildren, and great grandchildren of 201 and 200

represent Branch 1 of the BBF (the descendents of 409, 408 and 411, 410). Following the collection of Branch 1, Branches 2 and 3 were collected simultaneously. Branch 2 of the family, the descendents of 411 and 410, consists of eight siblings of 200, their children, grandchildren, and great grandchildren. According to family members, all of the siblings of 200 exhibited symptoms of mania and/or depression in their lifetime. Branch 3 of the family, the descendents of 409 and 408, is comprised of the siblings of 201, their children, grandchildren, and great grandchildren (Figure 2. 2).

Recruitment of Branch 1 was conducted in a systematic order; the proband helped recruit other family members beginning with first-degree relatives who in turn helped identify their relatives. In addition to the help of a number of especially motivated and co-operative family members in constructing the structure of BBF, genealogical records obtained from a local church in Senhora de Oliveira were available to confirm family relatedness dating back eight generations.

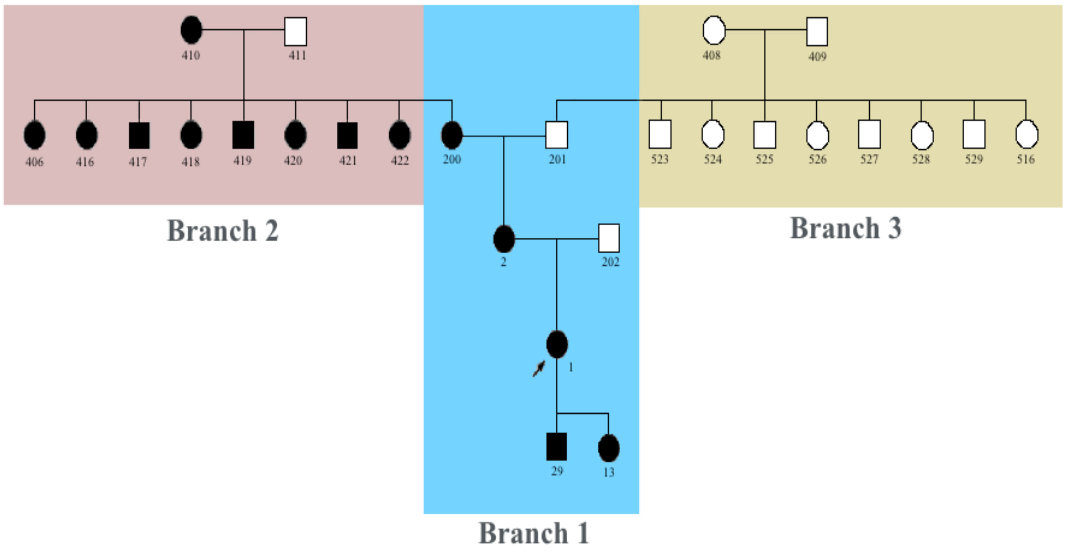


Figure 2.2. BBF branches ascertained through proband 1 indicated in the figure with an arrow. Branch 1 is related to Branch 2 through individual 201 and to Branch 3 through individual 200.

2.2 Ancestry Analysis of the BBF

The Brazilian population is known for its admixture, with a genetic background of three parental populations: European, African, and Brazilian Native Amerindians, and a wide range of diverse ancestry patterns in-between (Lins, Vieira, Abreu, Grattapaglia, & Pereira, 2010). The BBF members self reported mixed Southern European ancestry and this was confirmed by the interviewers. However, as genetic data was available and because admixture can be hard to identify at interview, analysis of the ancestry of the BBF was conducted using principal components, as implemented in EIGENSOFT version 3.0 (Price et al., 2006) *(see chapter 6 section 6.3.2.3)*. SNPs from fourteen BBF members, genotyped by deCODE genetics on the IlluminaHuman 610 QuadBeadChip (comprising 610,000 SNPs), were combined with SNP genotypes from the HapMap samples for individuals of Northern and Western European, Tuscan Italian, Japanese, Chinese, Mexican, Luhya, Massai, Yoruba and African American ancestry (Gibbs et al., 2003). The datasets were merged using PLINK version 1.07 (Purcell et al., 2007) after asymmetric and uncommon SNPs between the data sets were removed using a number of batch scripts. A total of 70,000 SNPs present in all population datasets and the BBF were used in the analysis. The principal components were plotted to visualise the BBF clustering with reference to the HapMap populations, which revealed that the BBF clustered more closely with the Northern and Western European and Tuscan Italian populations thus confirming the family's self reports of ancestry (Figure 2.3).

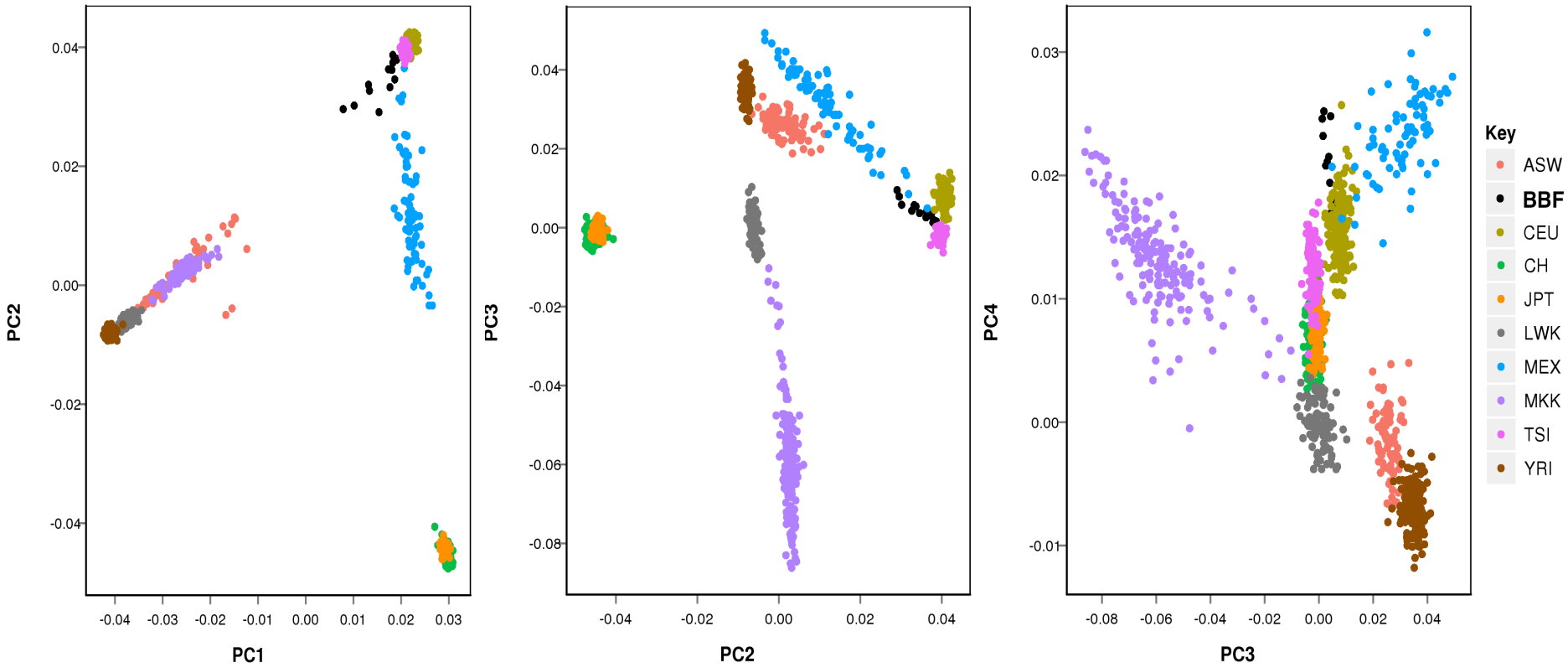


Figure 2.3. The graphs show the first four principal components (PC) from the EIGENSOFT ancestry analysis based on data from fourteen Brazilian Bipolar Family (BBF) members and ten HAPMAP populations. The first four principal components (PC1, PC2, PC3, and PC4) are plotted and show the BBF clustering more closely with the Northern and Western European (CEU) and Tuscan Italian (TSI) populations. Also shown in the graph are populations with African Ancestry from Southwest USA (ASW), Chinese from Beijing and Denver Colorado (CH), Japanese from Tokyo Japan (JPT), Luhya from Webuye Kenya (LWK), Mexican from Los Angeles, California (MEX), Massai from Kinyawa Kenya (MKK) and Yoruba from Ibadan, Nigeria (YRI).

Delineating the ancestry of the BBF was necessary to inform the linkage and case-control replication analyses that were going to be performed in the study and any conclusions that were going to be drawn from the findings of those studies in reference to published works. The analyses were not conducted using the Affymetrix 10K array data (comprising 10,000 SNPs) available for all BBF members because after asymmetric and uncommon SNPs between the BBF 10K data and the HapMap populations were removed an insufficient number of SNPs was available to perform the EIGENSOFT ancestry analysis. Pursuing the ancestry analysis further using alternative methods was not regarded as necessary given that analysis of the fourteen BBF members confirmed the reports of ancestry provided by the family and the interviewers.

2.3 Diagnostic Procedures

2.3.1 The Structured Clinical Interview for DSM-IV Axis I Disorders

All family members aged 17 years or older were given a face-to-face semi-structured interview, using the Portuguese version of the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID) (Del Ben, Rodrigues, & Zuardi, 1996) to systemically evaluate the presence of any current and/or lifetime Axis I psychiatric disorders. The SCID interview began with basic questions about demographics (e.g. age, marital status), educational history, and work history, followed by six sections on mood episodes, psychotic symptoms, differential diagnosis of psychotic disorders, differential diagnosis of mood disorders, substance-use disorders, anxiety disorders, somatoform disorders, eating disorders, and adjustment disorder. Participants were asked to report the presence or absence of DSM-IV diagnostic criteria with Yes or No

answers. The psychiatrists typically asked participants to elaborate on each answer with details of thoughts, feelings, and behaviours, where appropriate. Each criterion was rated as either absent or sub-clinically present with a negative (-), or as clinically present with a positive (+). Inadequate information for a positive or negative rating was coded with a question mark (?). Reaching a satisfactory response to SCID items sometimes entailed seeking corroborating information from other family members as well as challenging the participants' answers with more questions. On rare occasions, and given the tendency of bipolar patients to deny manic symptoms, the psychiatrists made clinical judgments regarding the presence of particular symptoms, even when the participants denied their presence.

2.3.2 The Kiddie-SADS-Present and Lifetime Version (K-SADS-PL)

Family members between the ages of 5 and 16 years old were administered the Portuguese version of Kiddie-SADS-Present and Lifetime Version (K-SADS-PL) (Brasil & Bordin, 2010), a semi-structured interview designed to assess current and past episodes of psychiatric disorders including major depression, bipolar disorder, dysthymia, cyclothymia, schizoaffective disorder, schizophrenia, attention-deficient hyperactivity disorder, and anxiety disorders, in children and adolescents according to DSM-IV criteria by interviewing the parent(s) and the child or adolescent. After information on the child or adolescent's general health, school functioning, peer and family relationships, and developmental milestones was collected, a screening interview surveying the primary symptoms of the different psychiatric disorders covered by the K-SADS-PL (e.g. increased goal directed activity, racing thoughts, decreased mood) was completed for current and most severe past episodes simultaneously. Each symptom was scored using either a zero to 3 point rating scale, where zero indicated no information was available, 1 indicated symptom was not

present, 2 indicated sub-threshold level of symptomology, and 3 indicated symptom was present and met criteria, or a zero to 2 point rating scale, where zero implied no information was available, 1 indicated the symptom was not present, and 2 indicated the symptom was present. Parent(s) and child reports were scored and a final summary rating for each item was achieved by including both sources of information. In cases of discrepant parent-child accounts, the interviewing psychiatrist used information available from other family members, as well as best clinical judgment to make a summary rating.

If the child or adolescent had clinical manifestations of symptoms associated with a particular diagnosis, a diagnostic supplement (either affective disorders, psychotic disorders, anxiety disorders, behavioural disorders, or substance disorders and other disorders supplement) appropriate to the symptoms displayed was administered as directed by the 'Skip out' criteria in the interview to aid the interviewer in making current and past DSM-IV diagnoses.

2.3.3 Scoring the Diagnostic Interviews

In total three hundred and eight interviews were completed by the BBF members, and only five eligible family members refused an interview. The interviewing psychiatrists and myself independently scored the diagnostic interviews and reviewed available clinical and hospital records before making independent diagnoses. Discrepancies in diagnoses were discussed extensively and reviewed by two independent psychiatrists who scored the diagnostic interviews in question and examined clinical records before making independent conclusions. A final diagnosis based on the consensus was used. Occasionally, additional patient information required to confirm diagnoses and was provided via telephone conversations with family members.

2.4 Diagnostic Profile of the BBF

A spectrum of mood disorders is manifested in the BBF with variable severity and duration, ranging from severe BPI cases requiring repeated hospitalisations to cases experiencing single depressive episodes with moderate severity. In total 111 (36%) interviewed family members received a mood disorder diagnosis. Forty (13%) family members fulfilled criteria for BPI, BPII, or SAD, which increased to 52 (16.9%) when bipolar spectrum disorders including BPNOS and cyclothymia were considered. Sex differences in the rates of mood disorders were observed in the BBF with a higher percentage of women (62.9%) reporting mood disorders than men (37.1%). These differences were maintained even when only BPI and SAD were considered as affected, with 61.3% women and 38.7% men reporting severe BPD episodes.

Unipolar depression segregated in the BBF with a high prevalence as 17.3% of family members received a diagnosis of major depressive disorder, or dysthymia. This lifetime prevalence, however, is difficult to interpret as estimates of lifetime prevalence of depression in the general population vary considerably between studies. For example the life time prevalence of depression was estimated to be 4.4% in the U.S multi-site Epidemiological Catchment Area Study (ECA) (Weissman et al., 1988), 8.56% in the general population of five European countries (Ayuso-Mateos et al., 2001) and 17.1% by the U.S National Co-morbidity Survey (Kessler et al., 1994).

Thirty-one percent of those affected with a mood disorder in the family exhibited co-morbid anxiety disorders, predominantly generalised anxiety and panic disorder. The prevalence of co-morbid anxiety did not, however, significantly differ from the prevalence of primary anxiety disorders reported in the family (13.2%) (χ^2= 0.01, p-value=0.93, df=1). In addition, alcohol abuse, defined as continued drinking

despite social, legal, or interpersonal problems, was reported by 21.6% of family members affected with mood disorders. This prevalence rate is above that reported in the general population, which ranges from 8.8 to 17.8% (Hasin, Stinson, Ogburn, & Grant, 2007). However, no significant differences existed between co-morbid alcohol abuse among family members affected with mood disorders and alcohol abuse among family members unaffected with mood disorders, 20.3% of whom reported alcohol abuse (χ^2= 1.0, p-value=0.32, df=1).

In addition to mood disorders, diagnostic procedures revealed one family member had schizophrenia, four family members were classed as mentally retarded, six suffered with adjustment disorder, two reported severe symptoms of premenstrual dysphonic disorder, five children were diagnosed with Attention Deficient Hyperactivity Disorder, and one child was thought to be on the autistic spectrum. In terms of physical disorders, the BBF had higher than expected rates of thyroidism (7.8%), type I diabetes (10.1%) and Parkinson's disease (2.6%).

2.4.1 Age of Onset of Mood Disorders in the Family

Reliable information regarding age of onset was available for 104 out of 111 affected family members. The average age of onset for all mood disorders was 29.6 years old (SD=±13.2) ranging from 7 to 85 years old (Figure 2.4).

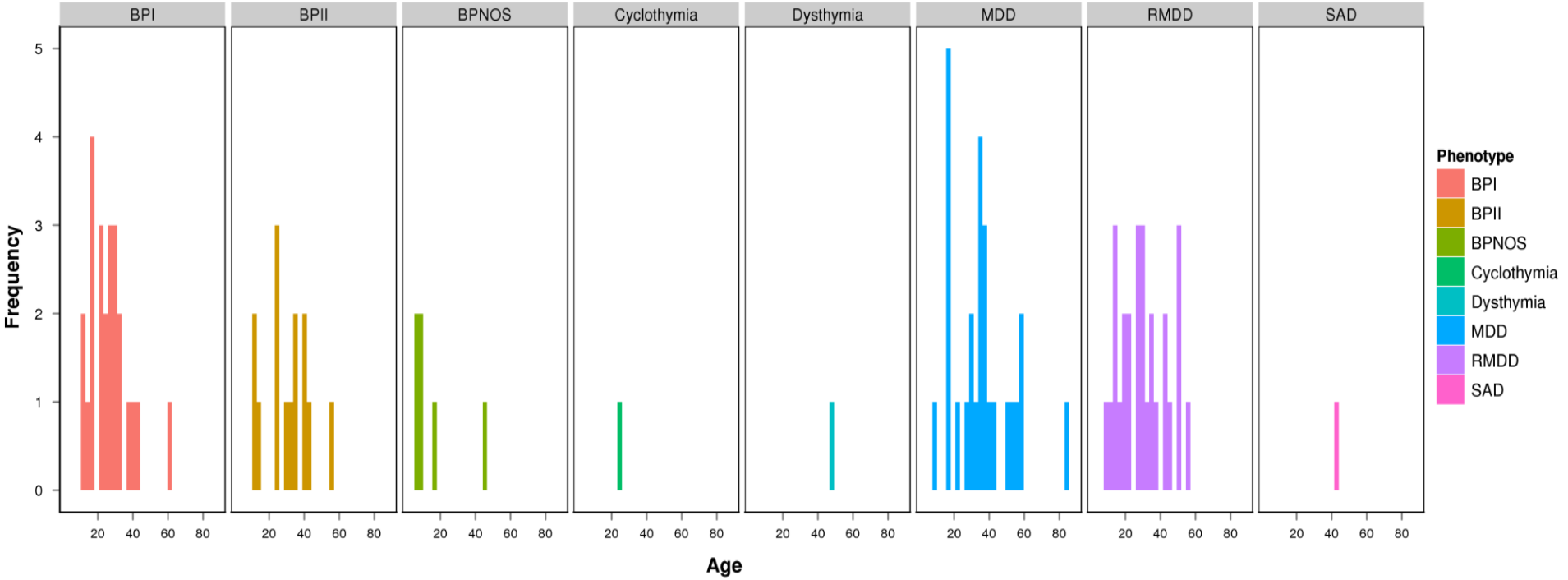


Figure 2.4. Histogram of age of onset by clinical diagnosis including bipolar I disorder (BPI), bipolar II disorder (II), bipolar disorder not otherwise specified (BPNOS), cyclothymia, dysthymia, major depression (MDD), recurrent major depression (RMDD) and schizoaffective disorder, bipolar type (SAD).

2.4.2 Paediatric Mood Disorders

In the BBF children as young as 7 years old received diagnoses of BPD (Figure 2.4). The research team was divided over such diagnoses for children under the age of twelve, mainly because identifying BPD in children is generally challenging and distinguishing BPD from Attention Deficit Hyperactivity Disorder (ADHD) and disruptive behaviour is sometimes difficult (Leibenluft & Rich, 2008). While adults have distinct periods of depression and mania that last for weeks or months, children have rapid cycling manic and depressive episodes that occur daily, sometimes simultaneously.

In general, paediatric BPD is the centre of debate among psychiatrists and British child psychiatrists tend to be more conservative with assigning paediatric BPD diagnoses than their American counterparts (Dubicka, Carlson, Vail, & Harrington, 2008; Post et al., 2008). However, some research suggests that paediatric BPD may not be rare, but only difficult to diagnose (Chang, 2009) and paediatric BPD might be a distinct subtype with a high familial loading (Bellivier et al., 2003). In the BBF, eleven children were under consideration for paediatric bipolar disorders. After consulting an experienced child psychiatrist, who reviewed the K-SADS-PL and interview notes, five children received diagnoses of BPNOS (two 7 year olds, one 9 year old, and one 10 year old), one received a diagnosis of BPII disorder (11 years old), and the remaining children were diagnosed with ADHD.

2.5 Phenotype Models

Bipolar disorder, SAD, and unipolar depression share common clinical features that are often indistinguishable. The similarities between the disorders extend to the genetic level, as family studies demonstrate their co-aggregation in families and molecular genetic studies suggest a "spectrum" of associated disorders with unique as well as overlapping genetic liabilities (Jones, Kent, & Craddock, 2002; McGuffin et al., 2003). These factors made deciding what should be classed as "affected" in the BBF difficult, especially that the nature and strength of genetic influences observed in a study may vary substantially depending on how precisely or how broadly the phenotype is defined (McGuffin & Sargeant, 1991).

For these reasons, we elected to use a hierarchical approach to affectedness. Three phenotypic models with widening groups of diagnoses classed as affected were constructed for the analyses: a narrow, broad, and super model, in addition to a depression only model. The narrow affection model included family members that fulfilled DSM-IV criteria for BPI, BPII, or SAD. The broad model included family members in the narrow model in addition to family members who fulfilled DSM-IV criteria for BPNOS and cyclothymia. The super model included family members in the broad model in addition to family members who experienced at least one episode of major depression of at least moderate severity (MDD) as defined by DSM-IV, or fulfilled diagnostic criteria for dysthymia. Finally, family members were included in the depression model if they had a history of dysthymia or experienced one episode of major depression (Table 2.1).

Diagnosis	Phenotype Model				Co-Morbid Disorders	
	Narrow	**Broad**	**Super**	**Depression**	**Psychosis**	**Anxiety**
BPI	24	24	24		19	10
BPII	15	15	15			3
SADB	1	1	1		1	
BPNOS		10	10		3	1
Cyclothymia		2	2			
RMDD			28	28	1	10
MDD			29	29		9
Dysthymia			2	2		1
Total (Percentage)	40 (13%)	52 (16.9%)	111 (36%)	59 (19.2%)	24 (21.6%)	34 (30.6%)
Mean Age Onset (SD)	27.7 (±11.2)	27.0 (±12.3)	29.6 (±13.2)	32.4 (±13.7)	N/A	N/A

Table 2.1. Phenotype models by number and percentage of affected family members and co-morbid psychosis and anxiety disorders. The total number of family members affected with a mood disorder is 111. The total number of interviews conducted is 308. The diagnoses are bipolar I disorder (BPI), bipolar II disorder (BPII), schizoaffective disorder, bipolar type (SAD), bipolar not otherwise specified (BPNOS), cyclothymia, recurrent major depressive disorder (RMDD) , one episode of major depressive disorder (MDD), and dysthymia.

Family members with mood disorders not fulfilling the particular diagnostic requirements for the phenotype model under study were labelled with affection status "unknown" in the analyses (as well as one schizophrenic individual, and seventeen very young children who provided blood samples but were not interviewed). Genotypes for individuals set as "unkonwn" are only used to infer missing parental genotypes and to provide haplotype phasing information (Terwilliger & Ott, 1994), which may increase the informativeness of the data. Individuals who were interviewed and did not receive any mood disorder diagnoses were labelled as "unaffected" in the analyses.

2.6 Anticipation in the BBF

The average age of onset of mood disorders significantly decreased in successive generations of the BBF with a mean age of 39.5 years old (SD±15.0) in the second generation compared with 31.7 years old (SD±11.3) in the third generation, 19.4 years old (SD±9.4) in the forth generation, and 14.0 years old (SD±3.6) in the fifth generation (Table 2.2). This suggests anticipation, a phenomenon characterised by earlier disease onset or increase in disease severity in successive generation, which was first described by Mott in 1910 in the offspring of "insane" parents being affected with disease earlier than their parents. Although anticipation has a molecular basis involving trinucleotide and other repeat expansion in certain disorders, apparent anticipation can occur for other reasons including ascertainment bias (Asherson et al., 1994; Penrose & Watson, 1945).

Anticipation was examined by comparing the age of onset in the second, third, fourth, and fifth generations of the pedigree. The first generation, most of whom are deceased, were excluded from the analysis due to insufficient information regarding age of onset. The differences in ages of onset were analysed with the ANOVA test. There were significant differences between the groups (p-value=0.013, F=1.87) suggesting anticipation was present. Assessment of disease severity across the generations was unfeasible as accurate measures of severity such as episode frequency were only available for a limited number of affected family members.

Diagnosis	Generation				
	I	II	III	IV	V
BPI		8	12	3	1
BPII		2	7	5	1
SADB					
BPNOS		1	4	6	
Cyclothymia			1		
RMDD		8	15	5	
MDD	1	3	17	7	1
Dysthymia			2		
Unaffected	1	26	86	88	17
Total	2	48	144	114	20
Age of onset Mean (SD)	85.0	39.5 (±15.0)	31.7 (±11.3)	19.4 (±9.4)	14.0 (±3.6)
Age of onset Range	N/A	15-65	13-56	7-40	11-18
Age at Interview Mean (SD)	93.50 (±12.0)	68.71 (±9.80)	44.64 (±14.85)	20.17 (±10.51)	10.10 (±7.26)

Table 2.2. Age of onset for affected BBF members across the five generations. The diagnoses are bipolar I disorder (BPI), bipolar II disorder (BPII), schizoaffective disorder, bipolar type (SAD), bipolar not otherwise specified (BPNOS), cyclothymia, recurrent major depressive disorder (RMDD) , one episode of major depressive disorder (MDD), and dysthymia.

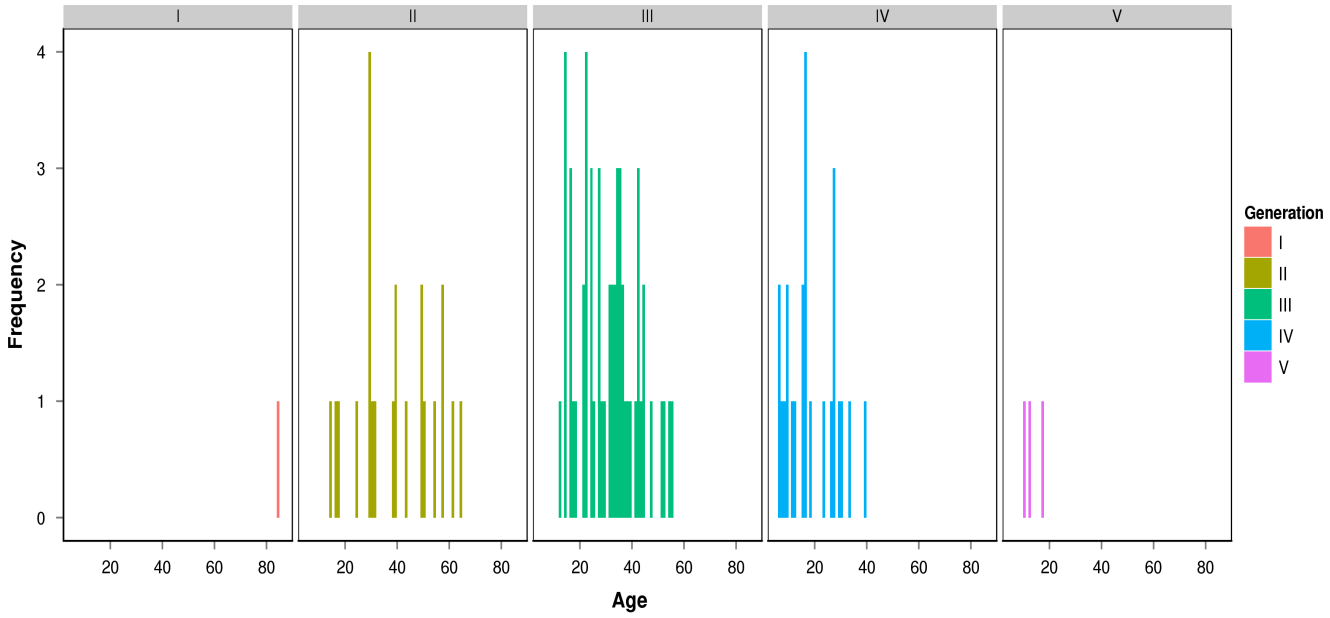


Figure 2.5. Age of onset distribution of mood disorders across the five generations of the BBF. Age of onset is graphed against its frequency in the family.

Table 2.2 and Figure 2.5 show the pattern of apparent anticipation. However, several sources of bias can mimic anticipation and need to be addressed before drawing definitive conclusions about the occurrence of anticipation in the BBF. Primary among these for the BFF may have been the differential availability of medical care for the younger generations versus the second generation who lived mostly in the isolated village of Senhora de Oliveira with sparse medical facilities, whereas later generations moved to nearby cities where psychiatric care was more readily available. As can be seen from Figure 2.5 some second generation family members received first time diagnoses aged 60 years or older, suggesting recall bias, as they may have forgotten the exact date their symptoms first emerged. Of note is that the clinical interview conducted in this study focused on the first appearance of mood episodes and most severe episodes encountered by family members rather than first episodes diagnosed. In addition, BBF members were known in their village for being "crazy and eccentric" and older family members described being weary of passing on "the madness" to their children. This awareness of disease could have lead to increased alertness in the parents and an earlier diagnosis in their children. The environmental stressors associated with an individual being raised by an affected parent could also have an effect on early onset of disease. Further, most family members in generations four and five were diagnosed and treated by our psychiatrists as a consequence of being interviewed for the study, and the BBF was ascertained through proband 1, whose child committed suicide aged 17 years during a manic episode. This introduces an ascertainment bias as described by Penrose & Watson (1945), who dismissed anticipation as an ascertainment artefact altogether.

2.7 Inbreeding in the BBF

The BBF is characterised by extensive inbreeding occurring over successive generations. Six instances of consanguinity exist between first-degree cousins; two in Branch 1, one in Branch 2, one between Branches 1 and 2, and two between Branches 1 and 3. In addition, two marriage loops in which siblings from unrelated families marry siblings from Branch 1 of the BBF exist. A section demonstrating the complexity of family relatedness in Branch 1 of the BBF is shown in Figure 2.6.

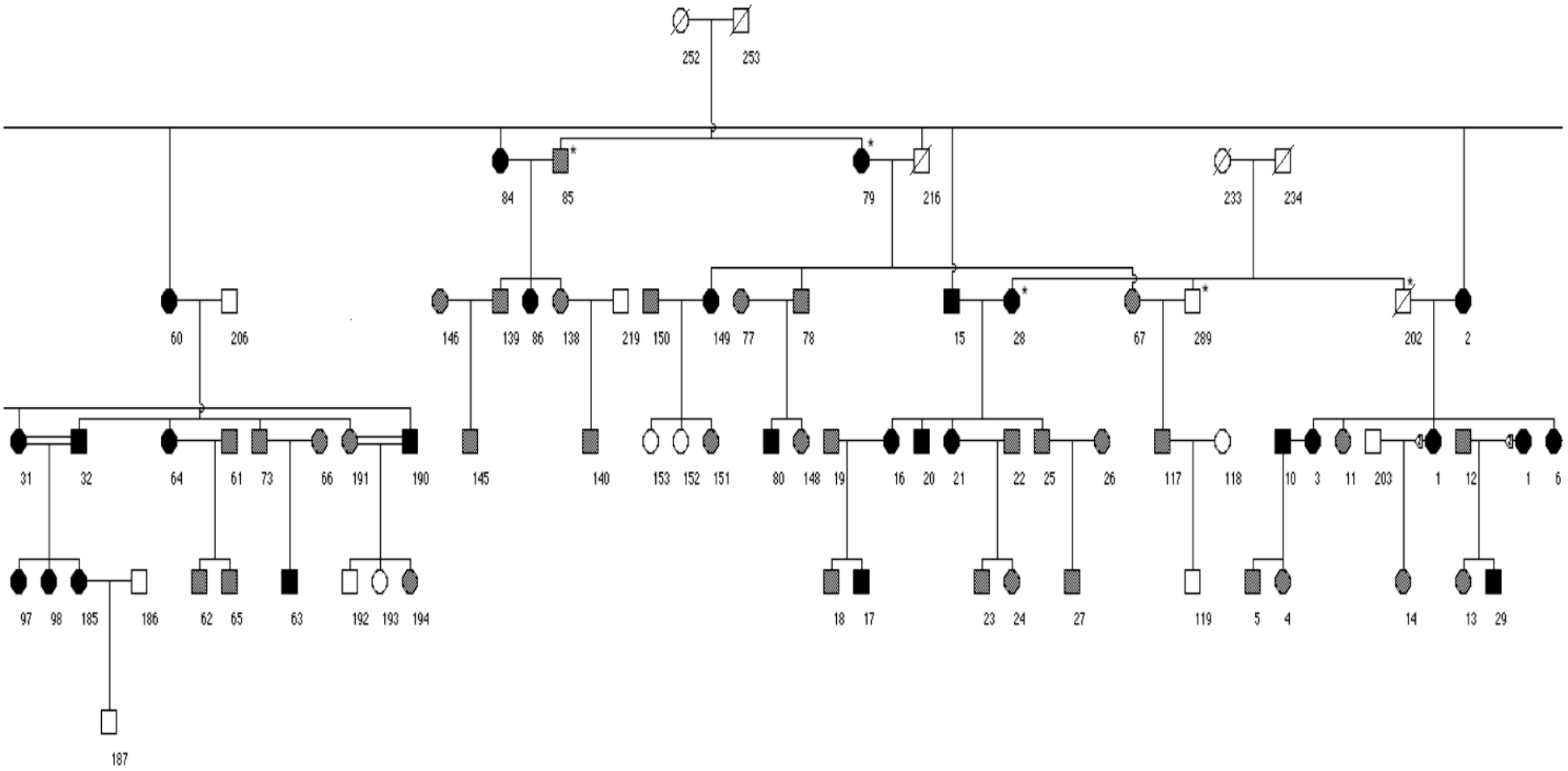


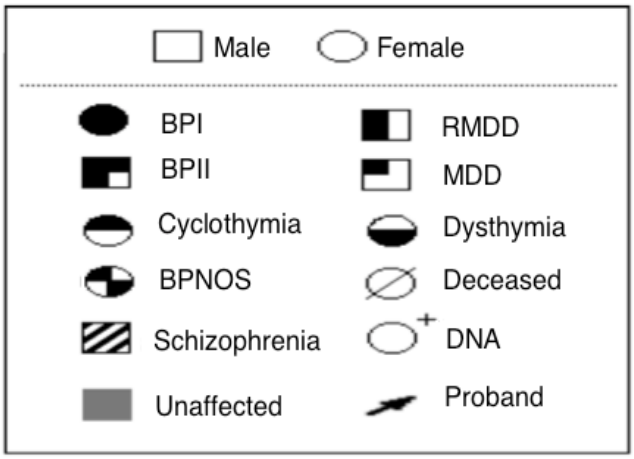
Figure 2.6. Consanguinity and marriage loops in Branch 1 of the BBF. Consanguinity loops are demarcated by two parallel lines and marriage loops are indicated with asterisks. Family members affected with a mood disorder are demarcated in black, unaffected in grey, and unknown in white. Symbols crossed by a line represent deceased individuals.

Reports on the health effects of inbreeding focus mainly on its impact on reproduction, childhood mortality, and rare Mendelian diseases. For a long time inbred families or isolated populations have been effective in detecting the causes of rare recessive diseases using homozygosity mapping, which hypothesises the presence in large inbred pedigrees, such as the BBF, of homozygous segments that are inherited by descent from a common ancestor. The idea being that homozygous affected individuals whose parents are related most likely receive a common haplotype without recombination from a single founder allowing linkage information to be gained from all presumed non-recombinant meioses all the way from the original founder haplotype to the affected children (Kruglyak & Lander, 1995). However, the usability of inbred families and isolated populations is extended to complex disease. Some studies have show that inbreeding has an effect on numerous multifactorial as well as Mendelian diseases and have emphasised the benefits of studying inbred families with common diseases. For example, a study found inbreeding caused an increase in homozygosity at many genetic loci with small deleterious effects on homeostatic pathways resulting in increased disease risk for blood pressure, unipolar depression, BPD, asthma, gout, and peptic ulcer (but not type 2 diabetes) in three genetic isolates from neighbouring islands in Middle Dalmatia, Croatia (Rudan et al., 2003).

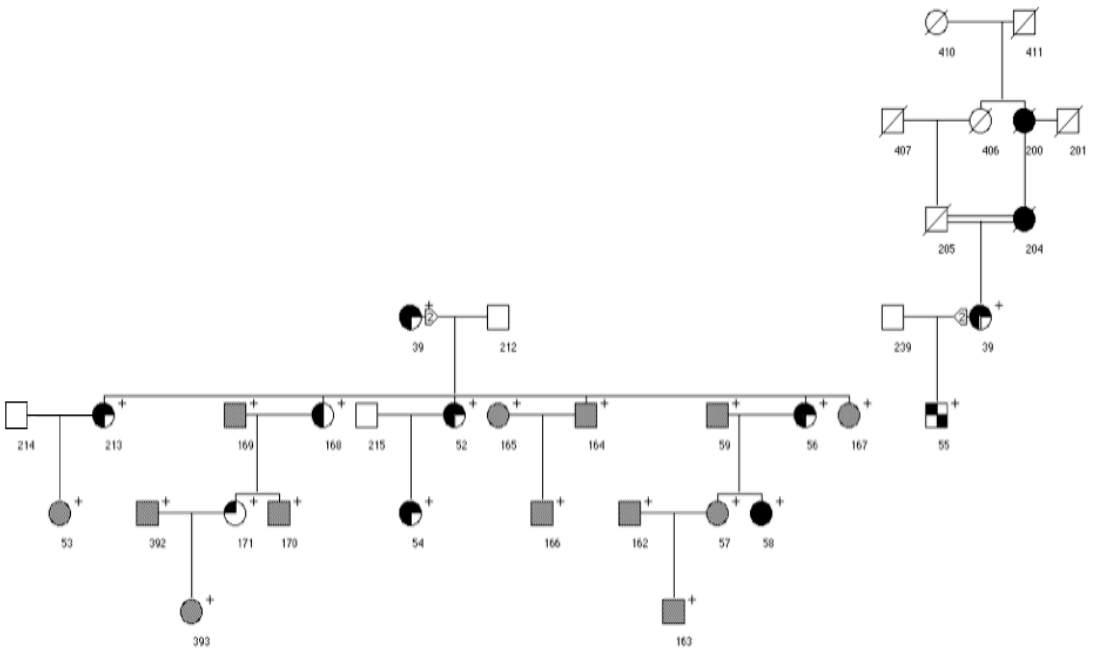
2.8 The Brazilian Bipolar Family Subfamilies

2.8.1 Branch 1 Subfamilies

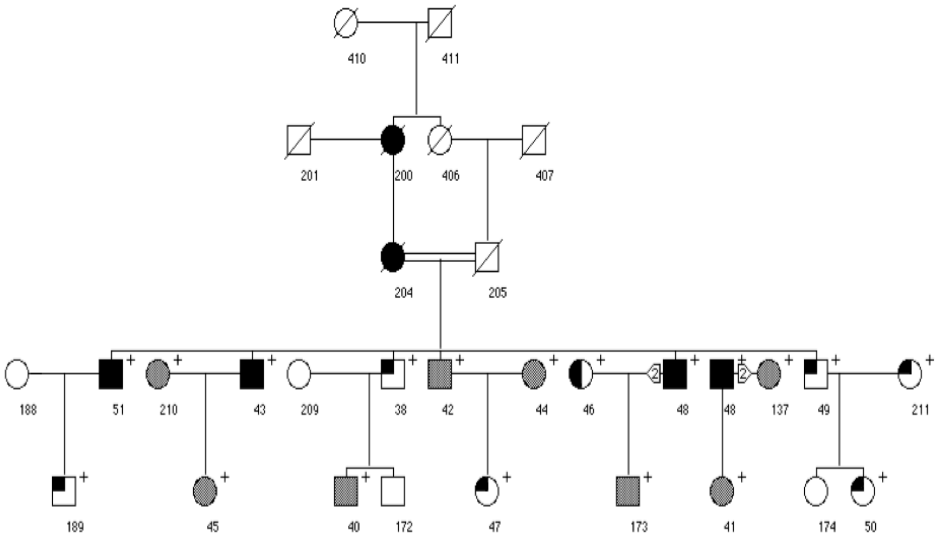
Branch 1 was divided into 12 subfamilies for the purpose of analysis as will be discussed later. This Branch of the BBF is densely affected with an array of mood disorders that vary in type and severity. A total of 214 individuals were interviewed from Branch 1. Assortative mating, or the tendency for individuals with similar phenotypes to mate more frequently than expected is common among people with mood disorders (Mathews & Reus, 2001) and occurs frequently in Branch 1. Assortative mating could partially account for the high incidence of consanguinity in the BBF. A pedigree diagram and description of each subfamily in Branch 1 is presented next.



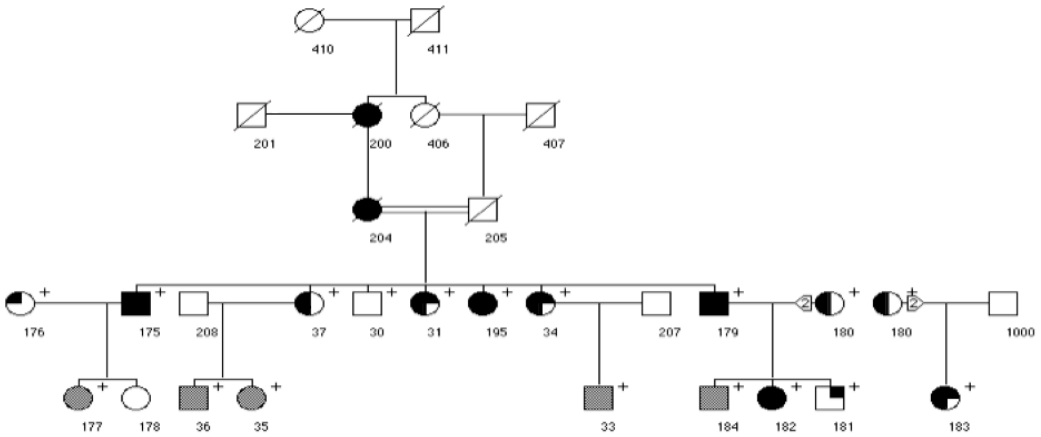
Key to Subfamilies 1-19. Eight disorders are included in the subfamily diagrams; bipolar I disorder (BPI), bipolar II disorder (BPII), cyclothymia, bipolar not otherwise specified (BPNOS), schizophrenia, recurrent major depressive disorder (RMDD), one episode of major depressive disorder (MDD), and dysthymia. Available DNA is demarcated with a plus (+).



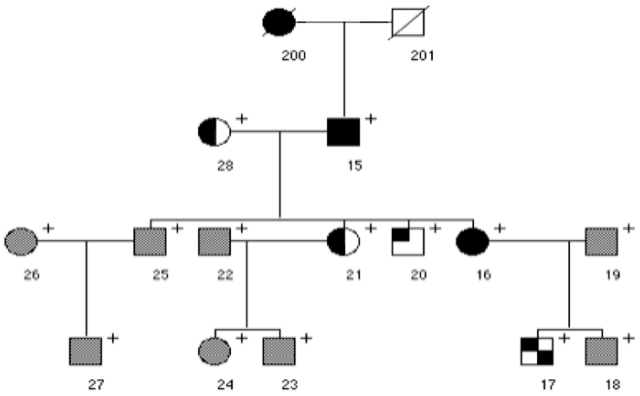
Subfamily 1. The parallel lines between 205 and 204 indicate a consanguineous marriage between two first cousins. 205 is a member of Branch 2 of the BBF and is suspected of being on the BPD spectrum and 204 had numerous hospitalisations due to puerperal mania. 205 and 204 are the parents of the largest sibship in the Branch 1 comprised of 16 children, presented in subfamilies 1 to 3. This subfamily predominantly segregates BPII disorder. 58 is the only member of the family with BPI disorder.



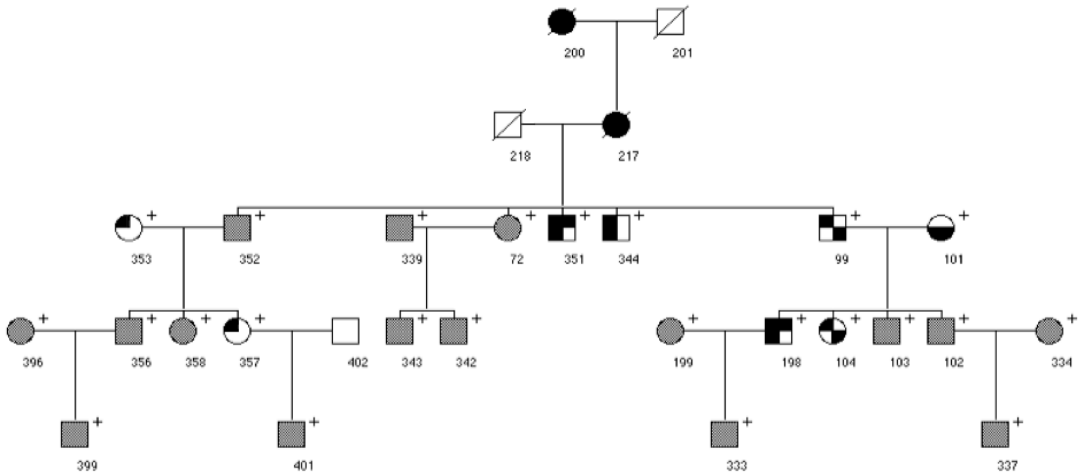
Subfamily 2. Continuation of the children of 204 and 205. All BPI cases in this subfamily were hospitalised due to psychotic mania. Individual 48 had severe BPI that necessitated over 15 hospitalisations and 10 altercations with the police to date. Individual 42 does not have an official SCID diagnosis, however, the interviewing psychiatrist noted he was hyperthymic during the interview.



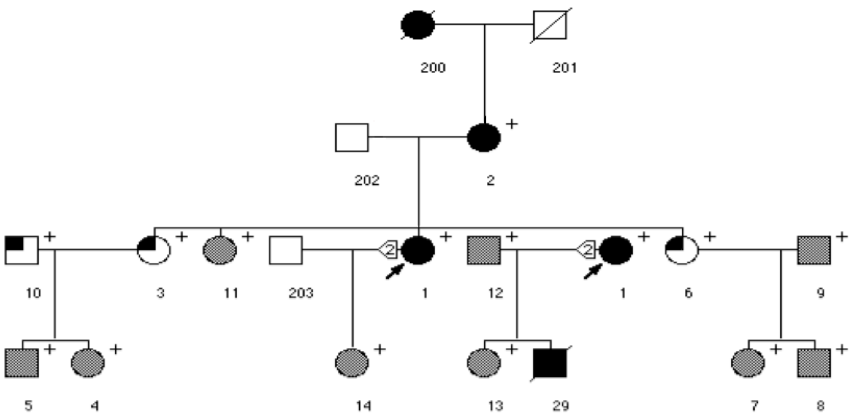
Subfamily 3. Continuation of the children of 204 and 205. Both individuals 175 and 179 were hospitalised on numerous occasions and reported visual and auditory hallucinations. Individual 179 had altercations with the law, as he tried to kill his wife, 180, on several occasions. Non-paternity was determined for individual 183. Individual 30 does not have an official SCID diagnosis, however, the interviewing psychiatrist noted he was hyperthymic during the interview.



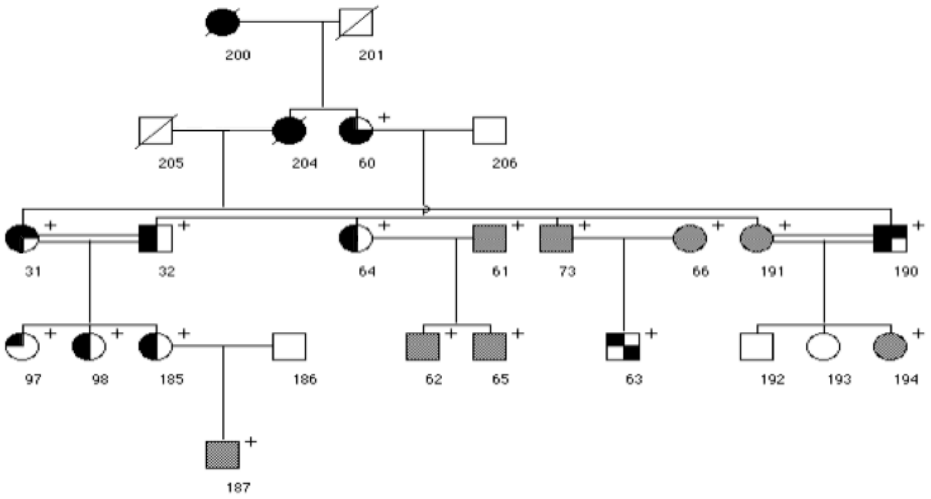
Subfamily 4. Individual 15 experienced a severe manic episode at 17 years old with psychotic symptoms, followed by numerous depressive episodes and currently only experiences hypomanic episodes (no medication). Individual 17 experienced antidepressant induced mania following a long period of depression at age 9.



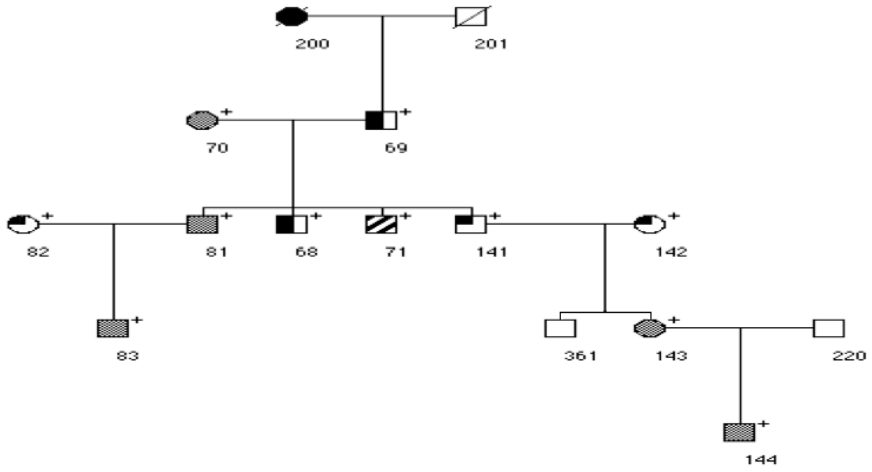
Subfamily 5. Individual 217 had a history of severe mania and died in a psychiatric hospital due to complications from diabetes. Individual 344 was receiving treatment for recurrent depression at the time of interview; however, psychiatrists suspect he might be on the bipolar spectrum (although he vehemently denied that). Individual 104 was diagnosed with BPNOS aged 10 years old.



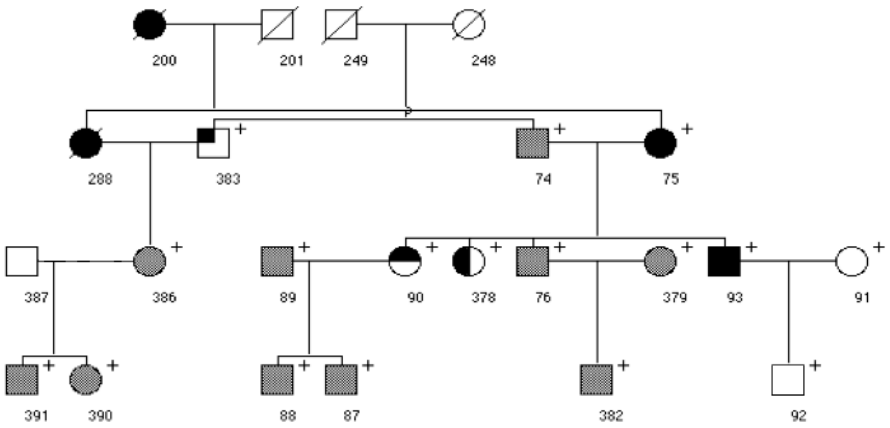
Subfamily 6. Individual 1 is the proband. Her manic episodes got worse after the suicide of her son, 29, who killed himself during a manic episode aged 17 years old. Individual 2 was hospitalised twice due to severe manic episodes and experienced over 20 episodes of debilitating depression.



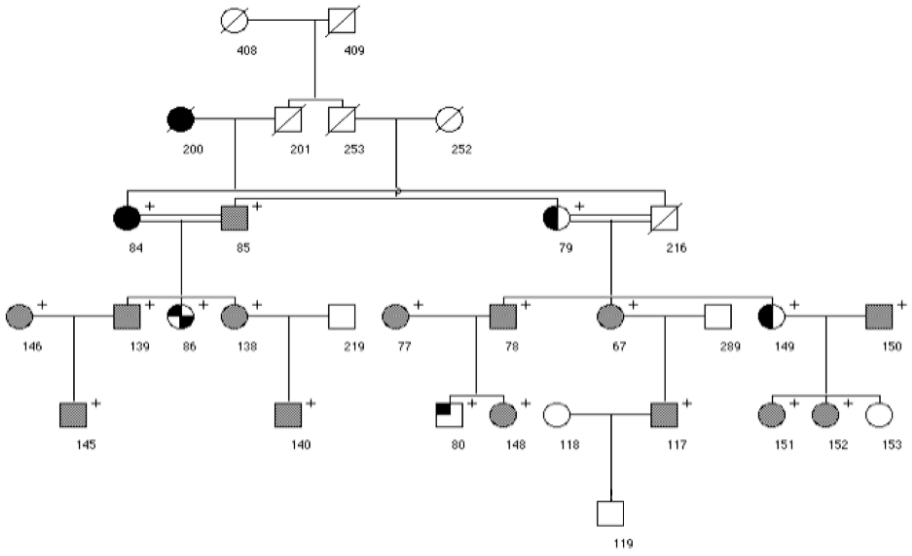
Subfamily 7. The parallel lines indicate consanguineous marriages between first cousins. Individual 63 is the youngest family member with a bipolar spectrum diagnosis. He presented with a mixed episode over a week during the psychiatrists' visit, oscillating between dysphoria, anger, impulsiveness and euphoria. Individual 98, aged 17 years old during the interview, reported multiple episodes of severe depression.



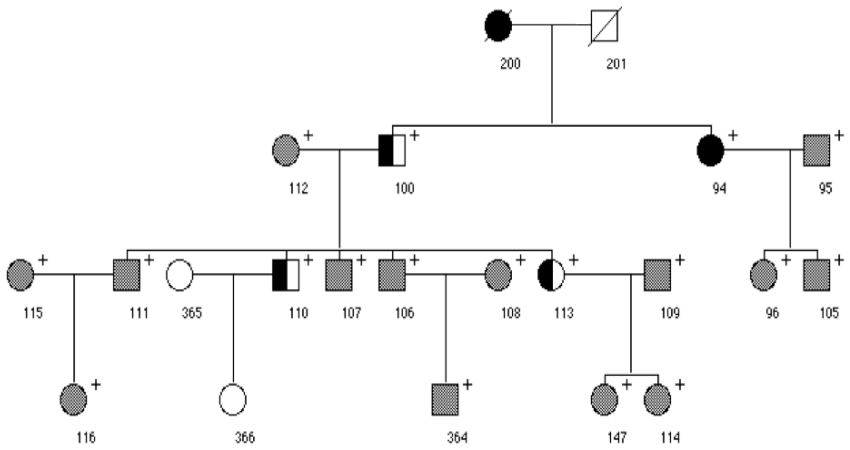
Subfamily 8. Individual 71 is the only BBF member with schizophrenia, paranoid subtype. He was diagnosed aged 37 years old. Individual 68 experienced multiple episodes of psychotic depression.



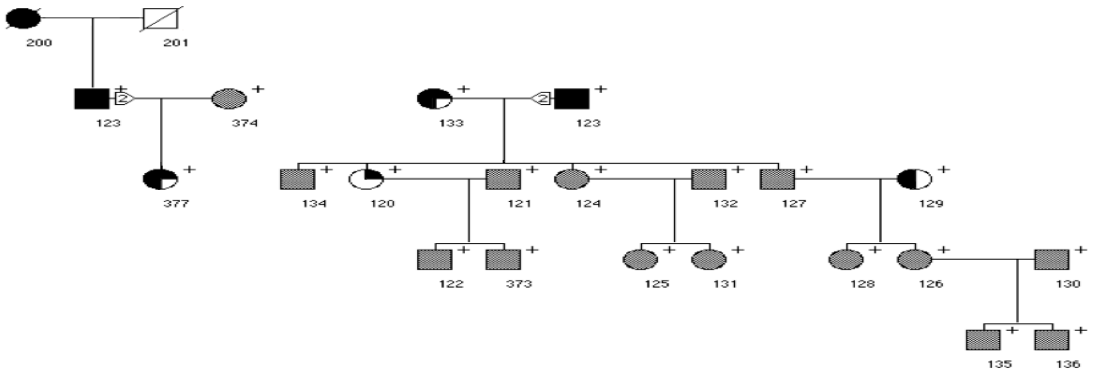
Subfamily 9. A marriage loop occurs in this subfamily, as two brothers (74 and 383) marry two sisters (75 and 288). Individual 75 experienced rapid cycling mania with psychosis. She built a church in her farm to practice her healing powers on town's people. Her symptoms increased after her son, 93, attempted suicide by climbing an electricity pole during a severe manic episode and lost complete function of one of his legs. Individual 93 was in a manic episode during the interview. Individual 92 is 10 years old. He is suspected of being on the bipolar spectrum based on information from his father. His mother denied he experienced any mood symptoms. He jumped out of the window to avoid meeting with the visiting psychiatrist.



Subfamily 10. The parallel lines indicate consanguineous marriages between first cousins. Individual 84 was hospitalised five times with severe manic episodes and psychosis. She has not experienced more episodes following treatment with Lithium. Individual 86 is deaf. Her mother reported 86 experienced periods of happiness and periods of crying and refusing to go to church that lasted no longer than a week, she received a diagnosis of BPNOS as the duration and severity of her symptoms was ambiguous.



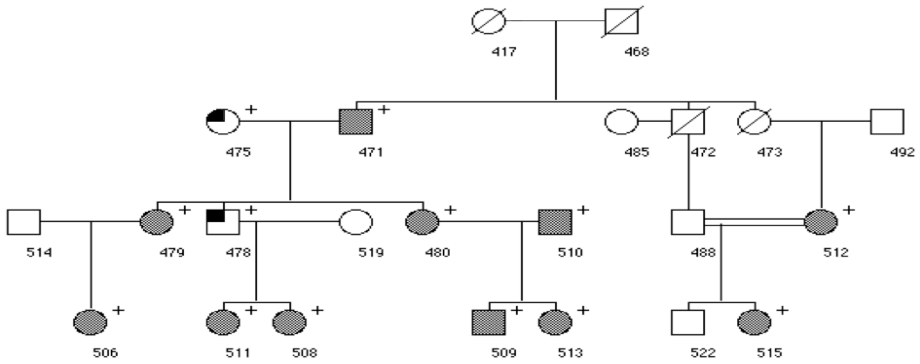
Subfamily 11. Individual 94 was hospitalised twice with severe psychotic mania and reported more than 10 episodes of depression. Her moods stabilised with Lithium. Individual 100 experienced multiple severe depressions and was hospitalised for a severe depressive episode for 20 days. He has been euthymic since starting on Lithium treatment. Individual 110 reported being diagnosed with depression aged 17 years old. Due to family history, he was prescribed Lithium by the study psychiatrists and his symptoms markedly improved.



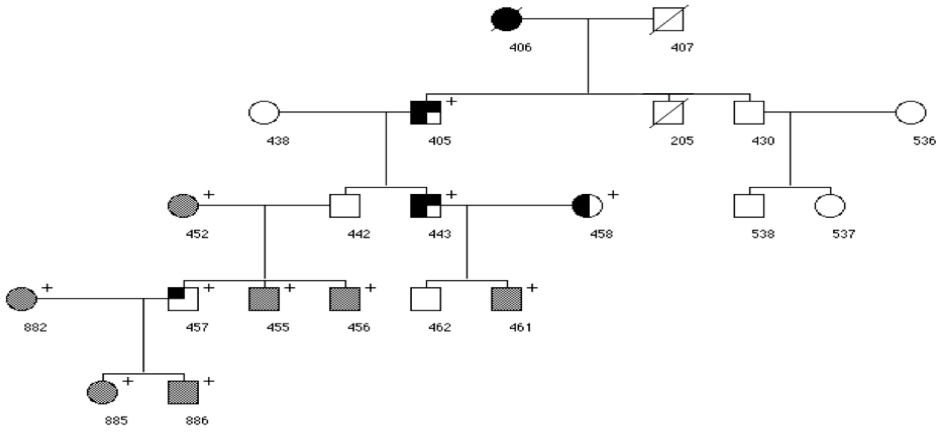
Subfamily 12. Individual 123 was hospitalised eight times with manic episodes and extreme aggressive behaviour. He inappropriately called his daughters to his bed when manic. His symptoms markedly improved with the introduction of Lithium. Individual 135 is a 6-year-old boy described by the visiting psychiatrist as odd and socially awkward. His mother is repeatedly called to school to address his impulsiveness and aggression.

2.8.2 Branch 2 Subfamilies

Branch 2 of the BBF is the least densely affected. None of the family members experienced severe symptoms warranting hospitalisation and none of them reported psychotic symptoms. Of note is that ascertainment of Branch 2 of the BBF was the least successful, as a large number of its members were either not contactable or refused participation in the study. Only 25 interviewed family members belong to this Branch. A pedigree diagram and description of each subfamily is presented next.



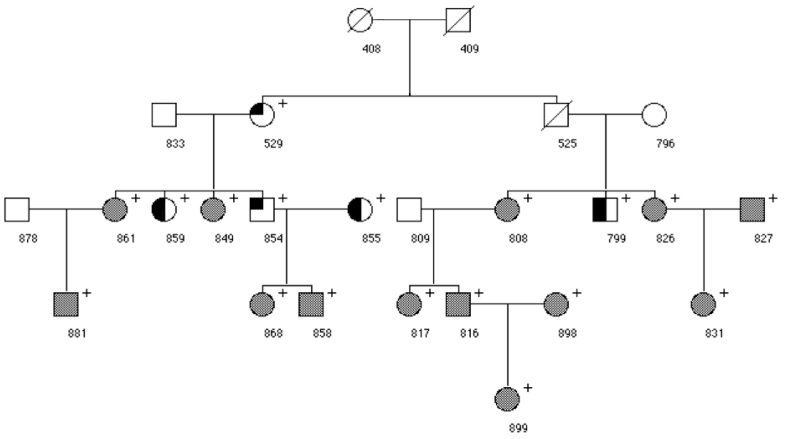
Subfamily 13. The parallel lines indicate consanguineous marriage between first cousins. This subfamily is not very densely affected, with two individuals, 475 and 468 reporting one episode of depression each. Individual 417 is suspected of being BPI based on information from family members and town's people.



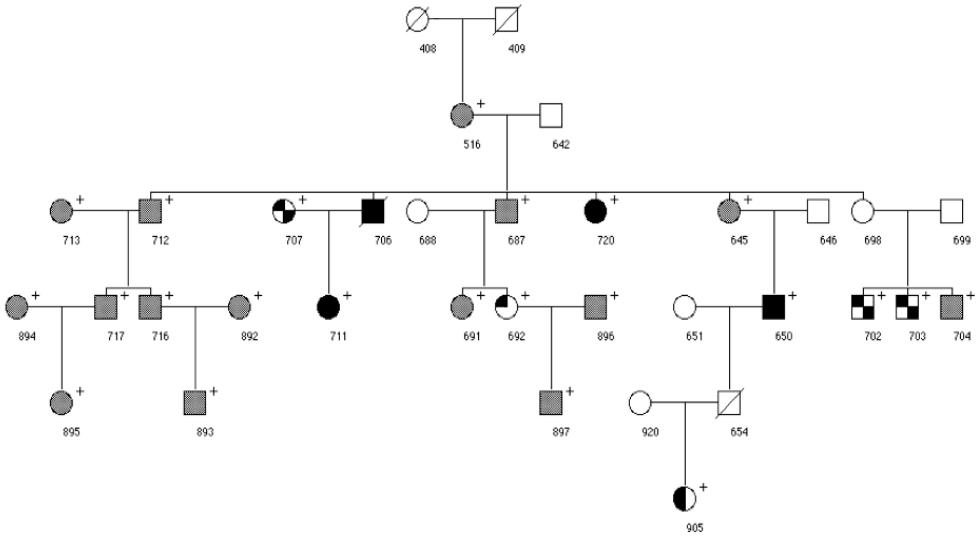
Subfamily 14. Individual 405 denied having any mood episodes, however, his son and grandson insisted he had periods of elated mood and renewed vigour as well as periods when he avoided family members and refrained from speaking. Individual 886 is suspected of being autistic.

2.8.3 Branch 3 Subfamilies

Branch 3 is the most out-bred of the BBF branches, with no consanguinity or marriage loops reported within the branch. In comparison to Branch 1, affected individuals in this branch do not exhibit psychotic symptoms. A total of 92 interviewed family member belong to this Branch. A pedigree diagram and description of each subfamily is presented next.

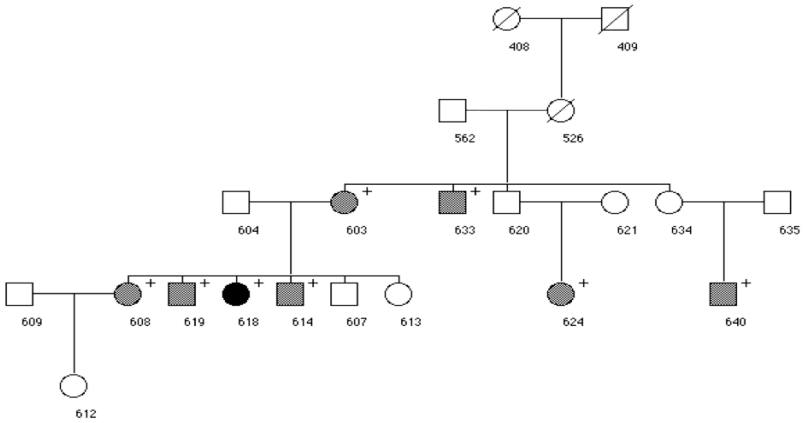


Subfamily 15. Individual 859 suffers from recurrent episodes of severe depression. Individual 799 experienced two episodes of depression, the first of which lead to voluntary hospitalisation for 90 days.

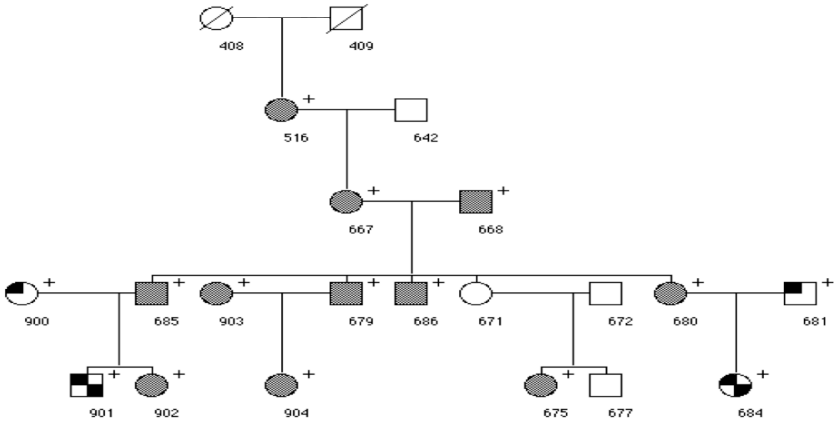


Subfamily 16. Individual 706 has hospital records indicating he suffered from BPI disorder. He committed suicide by shooting himself. Individual 711 experiences non-

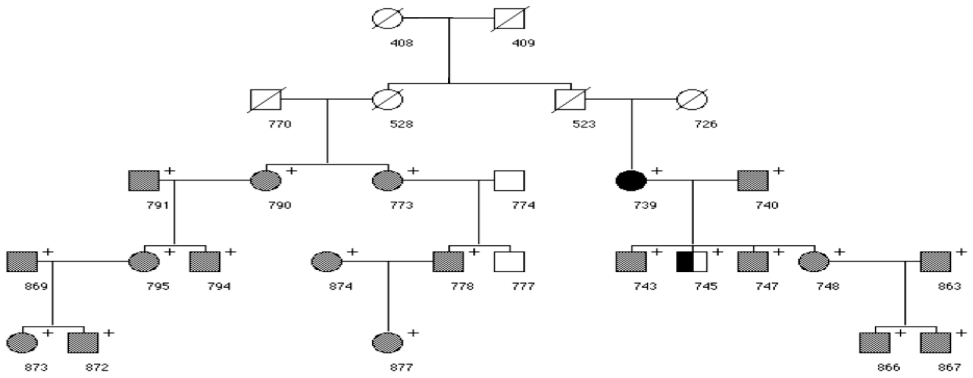
psychotic mixed episodes. Individual 654 committed suicide at the age of 22. Individuals 702 and 703 presented with manic episodes but no depressive episodes.



Subfamily 17. Individual 603 was described as hyperthymic during the interview. She revealed she had numerous extra-marital affairs and had 2 children with another man (not part of this study).



Subfamily 18. Individual 684 described short periods of irritability, agitation, and hypomania of moderate severity. Individual 681 was depressed during the interview, which was his first depressive episode aged 57 years old. Individual 901 received fulfilled K-SAD criteria for BP NOS, aged 7 years old.



Subfamily 19. Individual 739 was hospitalised five times after trying to kill himself during severe depressive episodes. Individual 745 had two episodes of depression that he received treatment for. The Diagnostic interview revealed he might have experienced a mixed episode, however, the interview was inconclusive.

Chapter 3 Linkage Introduction

3.1 Linkage and Recombination

In order to understand the principles of linkage analysis, it is necessary to appreciate how recombination occurs during meiosis. Each individual's genetic material is made of 23 homologous pairs one inherited from the mother and one inherited from the father, so that each gene is present in duplicate (except those on the X and Y chromosomes). During meiosis the homologous chromosomes crossover and exchange DNA segments in a process called recombination. In recombination a portion of the maternal homolog recombines with the paternal homolog to form a hybrid chromosome. This leads to the formation of gametes that possess non-recombinant and recombinant chromosomes that contain new combinations of alleles (Terwilliger & Ott, 1994).

Two genes that are on the same chromosome, transmitted from either the individual's father or the individual's mother, will be transmitted together unless a crossover point separates them. The closer the two genes are together, the less likely that a crossover point will occur between them and the more likely that they will be transmitted together (Sham & McGuffin, 2002). However, recombination events occur unevenly in the human genome. Certain "hotspots" are favoured sites of recombination, while areas in the vicinity of centromeres, are "recombination deserts", in which few corssovers ever take place. Typically, 80% of the recombination occurs in 10 to 20% of chromosomal sequence, with the exception of chromosome 19 which has a much lower density of hotspots (Myers et al. 2005).

Linkage is the main method of mapping disease loci in families. It tests the co-segregation between a disease locus and a genetic marker or markers such as microsatellites, multiallelic markers comprised of short, repeating sequences of DNA, or biallelic Single Nucleotide Polymorphisms (SNPs). Two methods of linkage analysis are used, parametric linkage and non-parametric linkage.

3.2 Parametric Linkage Analysis

Parametric linkage is a statistical method used to determine if a disease and marker locus co-segregate in families. Determining linkage would suggest that the disease locus is present in the region of the chromosome containing the marker. Parametric linkage analysis requires the specification of a disease transmission model, which details the mode of disease inheritance (e.g. dominant, recessive, or co-dominant), the disease allele frequency and the penetrance of the disease. Additional parameters such as marker allele frequencies, recombination rates between the marker loci, and the mutation rates at the marker loci are also specified, although the latter is often assumed to be zero (Schmidt, 2006).

3.2.1 The Recombination Fraction

The recombination fraction is the main measure of interest in parametric linkage analysis. It is the proportion of gametes that are recombinant with respect to two loci and is usually denoted as Theta (θ). It reflects the probability of recombination between two loci at meiosis. Two loci are said to be linked if during meiosis, recombination occurs between them with a probability of less than 50%. Conversely, two loci are said to be independently assorting, if during meiosis recombination occurs between them with a probability of 50%. The genetic map is based on the

recombination fraction between markers, where one Cenitmorgan (cM) corresponds to the occurrence of crossing over between two markers one time in a hundred. Delineating the recombination frequency between a disease locus and a maker with a known location on the genetic map allows an approximate localisation of the disease gene. Further, if additional markers in the region are examined then the disease gene can be narrowed down to a small region where the markers show no recombination with the disease locus (Sham and McGuffin, 2002).

3.2.2 The Logarithm of the Odds (LOD) Score

To evaluate the statistical significance of the recombination fraction (θ) between two loci, the logarithm of the odds (LOD) score is used. The LOD score compares the likelihood of the data under the assumption of linkage (θ is less than 0.5) to the likelihood of the data under the assumption of no linkage (θ is equal to 0.5) and could be expressed as follows:

$$LOD = Log_{10}\left(\frac{\theta^{R}(1-\theta)^{NR}}{(0.5)^{N}}\right)$$

Where *R* is the number of recombinant meioses, *NR* is the number of non-recombinant meioses and *N* is the total number of meioses.

If we consider a simple three-generation pedigree (Figure 3.1) where a double heterozygote (MmDd) father for a marker locus (Mm) and a dominant disease locus (Dd) mates with a double recessive homozygote (mmdd) mother for the same loci. Calculating the recombination fraction between the two loci is possible, and relatively straightforward, when phase, or the physical relationship between the two loci on the homologous chromosome(s) is known. If the two loci are co-inherited from the same

parent and are therefore physically present on the same homolog they are in coupling phase, with the opposite being repulsion phase (Ott, 1999).

By inspecting the illustrated pedigree we are able determine the linkage phase of the second-generation father, which is indicated by the vertical bar in his genotype separating maternal (left side of bar) from paternal alleles; D and M alleles are on one homologous chromosome and d and m are on the other. Both sets of alleles D M and d m are in coupling phase. Based on the inferred linkage phase we could determine with confidence that in the third generation five children are non-recombinant and two children are recombinant. The recombination fraction between the disease and marker locus is the total number of recombinant offspring divided by the total number of offspring, which is 2/7=0.29. Thus the gene or locus for disease D and marker M appear to be loosely linked with a recombination fraction of 0.29, suggesting that the disease and marker locus are approximately 29 cM apart.

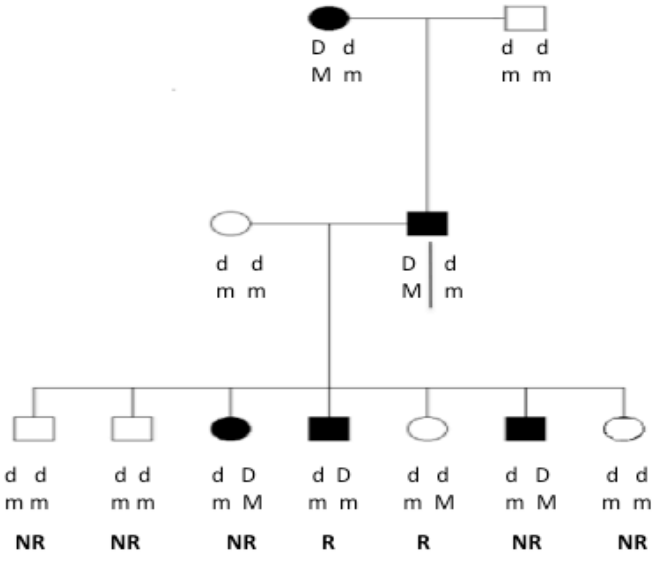


Figure 3.1 The father in generation II with DdMm genotypes is doubly heterozygous. It can be deduced that the alleles transmitted from the two grandparents to this parent are DM and dm. The 'coupling' of D and M and that of d and m, are known as the phase of the genotype. Five children are non-recombinant (NR) as they inherit the same genotype combinations as those transmitted in the previous generation and two children are recombinant (R) as they inherit genotypes that are unlike those transmitted in the previous generations.

To evaluate the statistical significance of this finding, the LOD score is calculated as follows:

$$LOD = Log_{10}\left(\frac{(0.29)^2(1-0.29)^5}{(0.5)^7}\right) = 0.28$$

By convention, a LOD score greater or equal to 3.0 is considered as significant evidence for linkage. It indicates 1000 to 1 odds that the linkage observed did not occur by chance. On the other hand, a LOD score less than -2.0 is considered as evidence to exclude linkage (Morton, 1955). Accordingly, the LOD score achieved from pedigree 3.1 neither confirms linkage between the disease and the marker locus, nor excludes it entirely.

In reality, phase is usually ambiguous due to incomplete pedigree information or inability to determine which two alleles were received by a double heterozygote parent from the maternal or the paternal gametes. Deducing the number of recombinant and non-recombinant offspring becomes difficult and calculating the LOD score becomes more arduous. To illustrate, if phase information for the double heterozygote father in pedigree 3.1 was unknown because genotypic information from the grandparents was unavailable or difficult to interpret because they are both heterozygotes, two outcomes from the segregating disease and marker loci would be considered depending on whether paternal alleles D and m were on one homologous chromosome and d and M on the other, or paternal alleles D and M were on one homologous chromosome and d and m on the other. In the first instance, and as demonstrated in pedigree 3.2 (a), we would have two non-recombinant and five recombinant offspring yielding a recombination fraction of 0.71, and in the second

instance we would have five non-recombinant and two recombinant offspring (b) yielding a recombination fraction of 0.29.

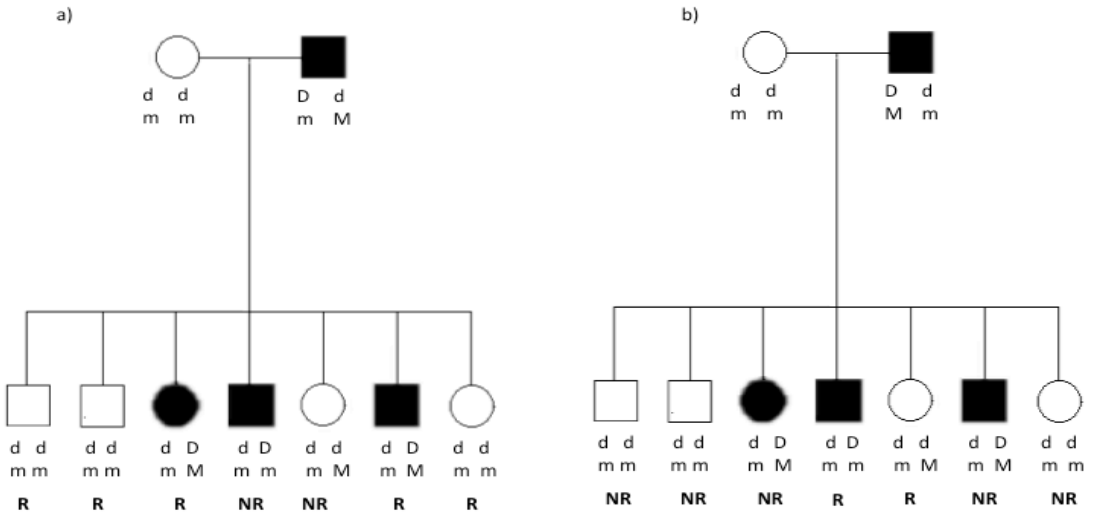


Figure 3.2. The father in generation II with DdMm genotypes is doubly heterozygous. Genotypes are not available for the grandparents, so the alleles transmitted from the two grandparents to this parent are of unknown phase. In a) two children are non-recombinant and four children are recombinant. In (b) five children are nonrecombinant and two children are recombinant. .

To calculate the LOD score the following formula would be used:

$$LOD = Log_{10}\frac{1}{2}\left(\frac{(0.29)^5(1-0.29)^2}{(0.5)^7}\right)+\left(\frac{(0.29)^2(1-0.29)^5}{(0.5)^7}\right)=0.02$$

Similar to pedigree 3.1, the LOD score achieved from pedigree 3.2 neither confirms linkage between the disease and the marker locus, nor excludes it entirely. This example demonstrates that with unknown phase calculating the LOD score becomes more difficult as all of the parent to offspring allele transmissions must be considered. The inclusion of extended families with missing genotypes, which is often the case in pedigree analyses, complicates the LOD score calculations even further as the number

of possible recombinant versus non-recombinant offspring combinations increases. In practice LOD score statistics are calculated for a range of recombination fractions.

3.2.3 Maximizing the Recombination Fraction

In parametric linkage a LOD score statistic is calculated for a range of recombination fractions (e.g. θ=0.0 to θ =0.3) reflecting every possible offspring outcome observed from the pedigree data. The recombination fraction achieved under the maximum LOD score reveals the best estimate of the recombination fraction, which is approximately proportional to the distance between the marker and disease locus. This is known as maximising the recombination fraction (Table 3.1). Tracing the recombination fraction between the disease locus and multiple loci with known locations in the genome allows the positioning of the disease gene or locus on the genetic map with some accuracy (Dawn Teare & Barrett, 2005).

Marker	Genetic Location	Recombination Fraction					
		0.0	**0.01**	**0.02**	**0.1**	**0.2**	**0.3**
1	40.0	2.94	2.23	2.21	1.49	1.10	0.73
2	44.0	2.71	2.70	**3.50**	1.18	1.50	0.90
3	48.0	2.30	**4.04**	2.90	1.54	1.44	0.65
4	50.0	2.93	2.90	1.90	0.98	0.90	0.59

Table 3.1. Example output from parametric linkage analysis using four markers with an inter marker distance of 4 cM at theta (θ) values of zero, 0.01, 0.02, 0.1, 0.2 and 0.3 indicating the disease and locus gene are in the same location, 1 cM apart, 2 cM apart, 10 cM apart, 19 cM[*] apart, and 27[*] cM apart respectively. The LOD score for marker 3 is maximised at a recombination fraction of 0.01 and the LOD score for marker 2 is maximised at a recombination fraction of 0.02 indicating that the disease locus is 1 cM away from marker 3 and 2 cM away from marker 2. Considering a LOD score of 3.0 as indicative of significant linkage, only markers 2 and 3 meet that threshold and therefore are used to localise the disease gene.

[*] Converted to genetic distance using the Kosambi function.
[*] Converted to genetic distance using the Kosambi function.

It is important to note that when the recombination fraction is small, the genetic distance and the recombination fraction are approximately equal. In practice geneticists treat them as equal for recombination fraction values of 0.1 or less. Recombination fraction values greater than 0.1, as demonstrated by the example in Table 3.1, are usually converted to genetic distance using either the Haldane or Kosambi function *(see Chapter 4 section 4.5.4.4)*.

3.2.4 The Disease Transmission Model

Parametric linkage was originally developed to map genes in simple Mendelian diseases with known genetic parameters (disease allele frequency, dominance relationship between normal and disease alleles, penetrance values) under a single locus two allele model defined by the formula:

$$K_p = f_1 p^2 + f_2 2pq + f_3 q^2$$

Where the population prevalence of the disease K_p is a function of the allele frequencies of the 'normal' allele and the 'disease' allele denoted by p and q respectively and three penetrance parameters denoted by f_1, f_2, and f_3 for the penetrance of zero, one, and two disease alleles respectively. Penetrance refers to the probability of being affected with a disease given a certain genotype. The penetrance set determines to what degree the phenotypic information of unaffected individuals is used for the linkage analysis (Terwilliger & Ott, 1994). The penetrance of f_1, also known as the phenocopy rate, is the percentage of sporadic cases in the population under study that are affected with the disease not due to genetic predisposition, but due to some unspecified environmental factor or to genes at other locations (not under study) (Ott, 1999).

Parametric linkage methods have been successful in identifying genes for dominantly inherited monogenic diseases such as Huntington's disease and recessively inherited diseases such as cystic fibrosis, but have been less successful when applied to genetically complex diseases.

3.2.5 Parametric Linkage Analyses in Complex Diseases

In complex diseases, such as BPD, where the mode of inheritance is not known assumptions about the disease allele frequency expected in the population; the disease penetrance, and the expected phenocopy rate are made. It is therefore common to analyse the data under several possible transmission models. This approach is called the maximized maximum LOD score (MMLS) (Greenberg, 1989), or mod score approach (Clerget-Darpoux, Bonaïti-Pellié, & Hochez, 1986). It typically involves analysing the data under a dominant and a recessive model and varying the penetrance parameters to include full penetrance, incomplete penetrance, or age-related penetrance. The robustness of this method has been verified by simulation studies, which show that maximising the LOD score over a range of transmission models, and subsequent correction for multiple testing, does not substantially decrease the power to detect linkage, compared with what one would find if the true mode of inheritance was used (Goldin & Weeks, 1993; Greenberg, Abreu, & Hodge, 1998). Authorities still contend that parametric linkage methods (as opposed to non-parametric methods) offer the highest power to detect linkage, when a genetic model that approximates the true mode of disease inheritance is specified (Abreu, Greenberg, & Hodge, 1999; Durner & Greenberg, 1992), and especially when the LOD score is maximised over several inheritance models.

3.2.6 Two-Point Parametric Linkage

The LOD score method described thus far is referred to as two-point linkage. It is carried out as a sequence of pair-wise comparisons calculating the recombination fraction between a disease locus and each of a number of markers. It is not affected by marker location anomalies, marker allele frequency misspecifications, or genotyping errors. Two-point linkage is, however, prone to error if the disease model is misspecified. Here misspecifying a dominant trait as a recessive trait or vice versa will strongly reduce the maximum LOD score and consequently the power to detect linkage. In addition, misspecifying the penetrance or disease allele frequency can result in overestimation of the recombination fraction where the LOD score maximum occurs, as non-penetrant cases may be misclassified as recombinants (Risch & Giuffra, 1992).

3.2.7 Multipoint Parametric Linkage

Estimating the position of the disease locus relative to a number of closely spaced markers with known location in the genome is known as multipoint linkage. The multipoint LOD score (MLOD) is defined as:

$$MLOD = Log_{10} \frac{Likelihood(data \mid x, \phi_0)}{Likelihood(data \mid x = \infty, \phi_0)}$$

Where ϕ_0 represents a specific genetic model, x represents the location of the disease locus on a specific marker map and $x=\infty$ refers to the null hypothesis of no linkage, i.e. the disease locus is not located on the marker map. Here linkage is usually expressed as a function of chromosomal positions measured in cM, as opposed to the recombination fraction in two-point methods. Multipoint linkage is in general more efficient than two-point linkage, as all relevant marker data are used simultaneously to

estimate the position of a disease susceptibility gene (Lander & Green, 1987). However, it is crucial that marker allele frequencies and the order of markers on the genetic map be correct as multipoint linkage simultaneously analyses several markers and therefore relies on the correct marker position and marker allele frequencies to locate the position of the disease locus on the genetic map. Errors in any of these parameters could result in misallocation of the disease susceptibility locus, or complete exclusion of linkage (Göring & Terwilliger, 2000).

3.2.8 Locus Heterogeneity and Heterogeneity LOD Score

Locus heterogeneity is the phenomenon of two or more independently acting loci each causing the same disease phenotype. Thus, if we perform linkage analysis with a marker located near one of these disease loci, some families will show linkage, whereas other families will show independent assortment of disease and marker and the combination of families will result in a reduction of the linkage signal, and may lead to failure to detect true linkage altogether (Abreu, Hodge, & Greenberg, 2002). To address locus heterogeneity, an admixture parameter alpha (α), which reflects the proportion of families linked to a particular locus can be incorporated into the LOD score analysis to allow for heterogeneity between investigated families, where an alpha of 1 indicates all members of a family show evidence of linkage to the locus under investigation, and an alpha of zero indicates that none of the family member show evidence of linkage. A heterogeneity LOD score (HLOD) is calculated thus allowing the detection of linkage even if some families under investigation do not have a genetic form linked to the locus of interest (Durner & Greenberg, 1992). The HLOD score assumes there are two categories of families in the data under investigation, some with $\theta = 0.5$ and some with θ less than 0.5, with a proportion α of the families segregating the disease gene ($\alpha > 0$) and a proportion of the families not

segregating the disease gene (α=0) and is calculated as follows:

$$HLOD = Log_{10} \frac{Likelihood(data \mid \theta < 0.5, \alpha > 0)}{Likelihood(data \mid \theta = 0.5, \alpha = 0)}$$

As with LOD scores the HLOD is maximised over disease models, and the maximum HLOD score achieved is taken to reflect the best estimate model of disease transmission. It is common practice in linkage to calculate both the LOD score and the HLOD score, and a more significant score under heterogeneity is taken to indicate locus heterogeneity.

In the investigation of large multigenerational families, within family heterogeneity might be possible. LOD scores achieved using the entire pedigree can be compared to LOD scores achieved when the pedigree is broken into smaller units. Splitting a large pedigree into smaller components reduces information, for example of phase and recombination events and therefore reduces power. However, if the disease allele is common, and intra-familial heterogeneity is present, LOD scores may actually increase when the pedigree is divided (Badner, Gershon, & Goldin, 1998). The proficiency of HLOD scores to accurately detect linkage has been shown to be in par with the that of LOD score analysis and a good deal of evidence indicates it is a powerful and robust tools for detection of linkage in the presence of heterogeneity (Hodge, Vieland, & Greenberg, 2002).

3.2.9 Allelic Heterogeneity

Allelic heterogeneity refers to the presence of different mutations within the same susceptibility locus so that different disease alleles at the same locus could each cause the same disease phenotype. An example of this is cystic fibrosis for which a number of different mutated alleles within the CF transmembrane conductance regulator gene

(CFTR) have been isolated and shown to contribute to the cystic fibrosis phenotype (Tsui, 1992). Contrary to the effect of locus heterogeneity on linkage, allelic heterogeneity does not impede linkage analyses (its effect is more pronounced in association methods of gene mapping).

3.3 Non-Parametric Linkage

As discussed, parametric linkage relies on the specification of a disease inheritance model and with the majority of complex diseases the underlying genetic model is difficult to estimate. As such, the use of non-parametric approaches has become commonplace in the linkage analysis of complex diseases and has proven to be successful in mapping disease loci for a number of complex diseases such as osteoarthritis (Hashimoto, Ochs, Komiya, & Lotz, 1998). The power to detect linkage using non-parametric methods has been reported to be equivalent to that of a LOD based method conducted under the correct inheritance model (Kruglyak, Daly, Reeve-Daly, & Lander, 1996).

The method is based on the idea that family members inherit the same alleles from a common ancestor at the disease locus (due to lack of recombination fraction between the disease and marker locus). Therefore, at a marker locus close to a disease-causing locus affected family members should share more marker alleles or haplotypes that are identical-by-descent (IBD), i.e. have the same DNA sequence and are inherited from a common ancestor, than would be expected under random Mendelian segregation. IBD is distinguished from identity-by-state (IBS), which occurs in unrelated as well as related individuals, and simply refers to the sharing of

genotypically identical alleles that are not inherited from the same ancestor. Of course if two alleles are IBS they are definitely IBD but the converse is not necessarily true. Assigning IBD status to affected family members is necessary for most NPL approaches (some rely on IBS). For example, consider the pedigrees in Figure 3.3, the siblings in pedigree (a) with marker genotypes 1/2 and 1/1 share allele 1 IBD (and IBS of course) because the latter child received a 1 allele from each parent one of which is IBD with the 1 allele the first child received. In pedigree (b) the siblings share two alleles IBS and one allele IBD as they must have received the same 1 allele from the father. There is no information about the IBD status of the maternally derived alleles because there is no way to differentiate the 1 alleles from each other. In pedigree (c) the siblings share zero alleles IBS and IBD and in pedigree (d) they share both alleles IBS and IBD. Pedigree (e) is more complicated in the sense that because both parents and children are heterozygous, each with genotype 1/2 we cannot tell whether there are two alleles IBD or none. If the children were homozygous (i.e. 1/1 each, 2/2 each or 1/1 and 2/2) then IBD status would be more easily determined as one allele IBD if they were both 1/1 or 2/2 and zero alleles IBD if they were 1/1 and 2/2. Determining IBD allele sharing is easier when all parental alleles are distinguishable from each other, due to extremely polymorphic markers, as in pedigrees c and d.

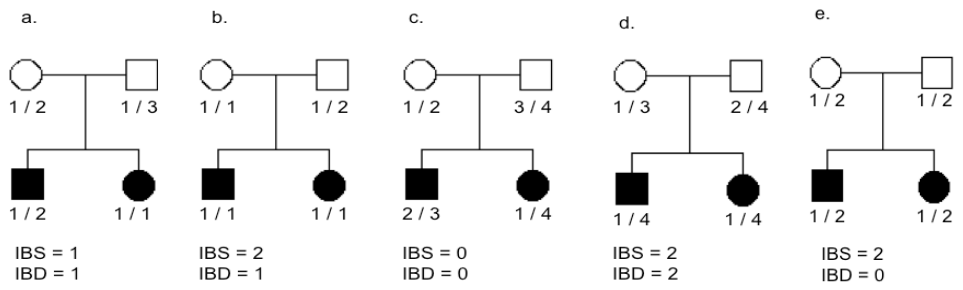


Figure 3.3. Affected sibling pairs with fully typed parents. Distinguishing Identical-By-State (IBS) from Identical-By-Descent (IBD) status in nuclear families with affected sibling pairs and fully typed parents. Distinguishing IBS from IBD is facilitated by the use of polymorphic makers. Adapted from Terwilliger and Ott (1994).

NPL approaches map disease susceptibility loci by looking for excess sharing of marker alleles or haplotypes among affected sibling pairs (ASP) and affected pedigree members (APM) (Terwilliger & Ott, 1994).

3.3.1 Affected Sibling Pairs

Originally non-parametric linkage methods were devised entirely for use in nuclear families with known parental genotypes and were most commonly based on affected sibling pairs (ASP). Here the probability of siblings sharing zero, one or two marker locus alleles that are IBD is ¼, ½ and ¼ under the assumption of no linkage. Hence the average number of alleles shared at a marker locus that is unlinked to the disease in ASP is ½, regardless of gene frequency or mode of disease inheritance (Sham & Zhao, 1998). The presence of linkage at a disease locus would lead to departure from this average and could be tested using the mean IBD test (Blackwelder & Elston, 1982), which evaluates the null hypothesis that the proportion of IBD allele sharing is equal to ½ by comparing the observed number of alleles shared IBD with the expected number. Assuming that IBD is known for certain and that in a sample of affected sibling pairs n_0 denotes sharing zero alleles IBD, n_1 denotes sharing one allele IBD, and n_2 denotes sharing two alleles IBD and N is the total number of affected sibling pairs, then the NPL test statistic could is defined as follows:

$$NPL = \frac{(2n_2 + n_1) - N}{\sqrt{(N/2)}}$$

In the ASP method the statistical analysis is simple as it only considers affected sibling pairs with available parental genotypes. The method becomes more complicated with extensions to include un-genotyped parents in the analyses or when

it is difficult to distinguish IBD from IBS (e.g. Figure 3.3 pedigree e) usually as a consequence of using markers that are not polymorphic enough to ensure that all parental alleles are distinguishable from each other (e.g. SNPs). In these instances NPL is determined using the prior and posterior probabilities of IBD sharing between sibling pairs. The prior probability is conditional only upon the relationship between the two siblings and is ¼, ½ and ¼ for sharing zero, one or two marker alleles that are IBD and the posterior probability is conditional on the genotype information only. For example, consider the pedigree in Figure 3.4 the mother is un-genotyped; she could have either genotype 1/3 or genotype 2/3. The posterior sharing probabilities for the siblings depend on the probabilities of these two maternal genotypes. If the mother is 1/3 the siblings share zero alleles IBD, however, if the mother is 2/3 they share one allele IBD. So we need to know the probability that a mother drawn form the population has genotype 1/3 versus 2/3. Using the Hardy Weinberg equilibrium where the probabilities (P) are P(1/3)=2P(1)P(3) and P(2/3)=2P(2)P(3) would determine the most likely maternal genotype. Here the posterior probability depends on the assumed allele frequencies (Cordell, 2004).

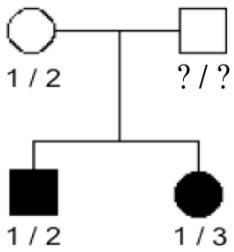


Figure 3.4. Nuclear family with unknown maternal genotype. The maternal genotype must be either 1/3 or 2/3, which depends on the population genotype frequencies assumed for the maker frequencies in the analysis.

To calculate the NPL, a maximum Likelihood score (MLS) could be calculated and is defined as follows:

$$MLS = Log_{10} \frac{Likelihood(z_0, z_1, z_2)}{Likelihood(0.25, 0.5, 0.25)}$$

Where Z_0, Z_1 Z_2 represent the maximum likelihood estimates of the posterior probability of zero alleles IBD, one allele IBD, and two alleles IBD over the likelihood of the prior probability (Cordell, 2004).

3.3.2 The Affected Pedigree Member Method

The Affected Pedigree Member (APM) method considers all affected family members in nuclear and extended pedigrees. Whittemore and Halpern (1994) introduced two non-parametric linkage statistics, NPL_{pairs} and NPL_{all} that are commonly used in extended pedigrees. The tests are based on the posterior probability of IBD allele sharing (based on the genotype data given the family relationship) compared with the expected prior probability of IBD sharing (based on the individuals relationship only). Observed IBD allele configurations for affected family members are assigned scores on the basis of family relationships and frequency of the marker alleles in the population. The scores for all IBD configurations possible in the pedigree are summed and a normalised score is calculated under the assumption of no linkage. The NPL test for pedigree *i* could be defined as

$$\overline{Z}_1 = \frac{(\overline{S}i - \mu_i)}{\sigma_i}$$

Where *S* is the sum of IBD configuration scores, μ_i and σ_i are the mean and standard deviation of all possible IBD configurations calculated on the basis of family relationships regardless of whether they are compatible with the observed genotypes. NPL_{pairs} tests compute IBD scores for pairs of affected family members, while NPL_{all}

tests compute scores for arbitrary groups of affected family members. The NPL_{all} test is complicated given that the number of possible IBD configurations grows rapidly with number of individuals considered. For example, three IBD configurations are possible in two individuals, sixteen in three individuals, and 624,889 in seven individuals (Whittemore and Halpern, 1994).

3.3.3 Are Non-parametric Linkage Approaches Truly 'Model Free'?

NPL methods that depend on the specification of marker allele frequencies to calculate the probabilities of IBD sharing among family members, such as NPL_{pairs} and NPL_{all} are not entirely 'model free.' They are only model free in the sense that they do not rely explicitly on disease model parameter specifications, i.e. estimations of penetrance, phenocopy rate, and mode of disease transmission are not required for the analysis. However, they have implicit model assumptions inherent in the scoring functions and weighting parameters employed by the NPL test statistic used (Kong and Cox, 1997; Farrall, 1997; Kruglyak, 1997) and rely as discussed on the population marker allele frequencies specified in the analyses.

NPL_{pairs} and NPL_{all} have been the topic of debate in terms of the power they possess to detect linkage in single-locus, multi-locus, nuclear and multigenerational families. The consensus, based on multiple simulations comparing different ways of scoring allele sharing between affected relatives, holds that NPL_{all} tests tend to perform better and therefore detect linkage when the underlying mode of disease transmission is dominant (Kruglyak, 1997), while NPL_{pairs} tests tend to perform better, and therefore detect linkage, when the underlying mode of disease transmission is recessive. The jury is still out with regard to diseases with additive genetic inheritance, as some statistical geneticists attribute better power to detect

linkage when the underlying genetic model is additive to NPL_{all} test statistics, while other attribute it to NPL_{pairs} (Whittemore and Halpern, 1994). Of note is that all of the assumptions made regarding the power of different NPL tests, depend to a large extend on the size of the simulated data, the type of family under study (nuclear vs. multigenerational) and the density of the genetic map used and are therefore not necessarily applicable to all datasets.

3.3.4 Affected Only NPL Designs

NPL statistics are usually conducted using genotypes from affected family members only. The rationale behind this is that affected family members contribute most of the linkage information and elimination of unaffected family members does not usually cause a severe loss of power. Moreover, in more complex diseases individuals classified as unaffected may eventually develop the disease at a later stage making the designation of a family member as 'unaffected' more uncertain than the designation of a family member as 'affected.' Therefore looking only at the phenotype of affected individuals avoids mistakenly assigning a low risk genotype to individuals in whom the mutation in not yet penetrant (McPeek, 1999). Even though NPL tests score affecteds only, genotype information from unaffected family members, particularly siblings, is used to infer missing parental genotypes, thus making inferences about IBD sharing more precise.

3.4 Whole Genome Linkage Significance

A LOD score of 3.0 (p-value=0.0001) is conventionally used to indicate significant linkage and denotes the likelihood of the observed pedigree data (including information on pedigree structure, phenotypes, and marker genotypes) is 1000 times

higher under linkage than under independent assortment, and a LOD score of -2 is used to indicate significant evidence against linkage (Morton, 1955). This convention was challenged by Lander and Kruglyak (1995) who contended to keep the chance of encountering a false positive at no more than 5% or once in 20 genome scans, one must impose a more stringent LOD score of 3.3 (p-value = 4.9×10^{-5}) for whole-genome linkage and a LOD score of 1.86 (p-value=1.7×10^{-3}) as suggestive of linkage. The stringency of the proposed significance levels are the source of constant debate in the literature since they are based on the assumption of a dense marker map with no missing data.

3.5 Linkage Studies of Large Multi-Generational Families

When the genetic component of a disease is caused by multiple distinct genes/loci with major effect, the power to detect linkage is reduced due to locus heterogeneity. Studying large multigenerational families has been proposed as a means of reducing heterogeneity, as they are more genetically homogeneous and probably segregate fewer disease-causing genes/loci than a collection of independent nuclear families. The caveat is, however, that in common diseases, extended families may feature intra-familial heterogeneity, proportional to the size of the pedigree, caused by genes introduced by marrying in-spouses. In this case, the loss of power to detect linkage due to heterogeneity is compensated by the increase in genetic information available for analysis and heterogeneity must be accounted for in appropriate test statistics. Dividing a pedigree into its nuclear constituents would be beneficial in this situation. Locus heterogeneity could also be attributed to loosely defined disease phenotypes or the use of multiple subtypes of a disease that might themselves be diseases with distinct causes, each with a distinct genetic component. Because large families usually segregate more severe and clinically similar forms of a disease, they are considered

ideal candidates for linkage studies (as well as isolated populations). In addition, penetrance may be more appropriately defined in a large pedigree where information on anticipation, imprinting, and other factors are likely to be more available (Blackwood & Muir, 2001) and because the probability that a major disease gene is segregating in a large family is increased, affected individuals have a lower probability of being phenocopies compared to affected individuals in the general population (McInnes et al., 1996). Large families are also more environmentally homogenous, and offer more genetic information that enables estimation of gametic phase for un-genotyped family members.

On the flip side, large families are extremely difficult and expensive to collect and come with analytic difficulties inherit in their size and complicated structures. Enormous computational difficulties are reported by most large family linkage studies that resort to breaking large pedigrees into smaller, more "manageable" units for analysis.

3.6 Genetic Markers in Linkage Studies

In order to be useful for mapping disease genes, marker loci must be highly polymorphic so their segregation in families can be tracked accurately. When a parent passes a locus to an offspring determining whether it is recombinant or non-recombinant depends to a large extent on whether the parent is a double heterozygote, i.e. heterozygous at each of the two loci under study. Similarly, distinguishing IBD from IBS allele sharing between siblings or affected family members is more easily and accurately accomplished when the investigated markers are highly polymorphic. For these reasons microsatellite markers were used in linkage studies for more than two decades. The high polymorphism rates of these loci and low cost PCR procedures

made them attractive. However, more recently, SNPs gained popularity as genetic markers in linkage studies (Figure 3.5).

Figure 3.5. Microsatellites are multiallelic markers comprised of short, repeating sequences of DNA. The repeat motif is typically two to four base pairs long. The alleles of a microsatellite markers differ from each other in the number of repeat units. A tetranucleotide is shown in the graph with 2, 3, and 4 repeats. SNPs are biallelic and comprise changes in a single base pair. Mapping disease genes using microsatellites lost favour to SNPs only due to the ability to genotype SNPs at large volumes.

Although biallelic and therefore less polymorphic than microsatellite markers, the high abundance of SNPs, occurring every 100 to 300 base pairs along the 3-billion-base human genome, coupled with the advent of highly automated, cost effective genotyping platforms, meant that their low informativeness could be offset by the ability to genotype them in large numbers. It is reported that a marker density of one SNP every 4.5 cM, which requires approximately 700 to 900 polymorphic SNP markers for whole genome coverage, has the same information content as approximately one microsatellite every 10 cM, which requires between 300 and 400 microsatellite markers for whole genome coverage. This substitution is, however, only workable if the minor allele frequency of the SNPs used is 20% or above

(Kruglyak, 1997). Studies comparing the performance of SNP arrays to microsatellites have shown that current SNP genotyping arrays (e.g. Affymetrix 10K array) offer higher genomic coverage, have greater information content (due to large numbers), and produce lower genotyping errors than microsatellite markers. Several studies that have re-analysed existing microsatellite linkage scans with denser maps of SNPs have found either suggestive or significant linkage missed by the initial scans (John et al., 2004; Middleton et al., 2004). The variability in linkage data due to use of microsatellites versus SNP markers is a contributor to the inconsistency of linkage findings due to differences between the two genetic maps.

An issue arising from SNP maps concerns LD, as dense SNP maps are more likely to contain markers with strong LD than microsatellite maps. Given that linkage is conducted under the assumption of linkage equilibrium, LD could potentially be problematic in multipoint linkage analyses as unaccounted for LD may increase false positive rates especially when parental genotypes are missing (Scott, Schmidt, Ashley-Koch, & Schmidt, 2005) as haplotype frequencies can be incorrectly inferred. LD could be accounted for by correcting LOD scores for inflation due to tight LD between SNPs or alternatively SNPs with high LD could be treated as one single marker in the analyses (Abecasis & Wigginton, 2005).

Chapter 4 Materials and Methods

4.1 Sample Collection

Following the diagnostic interviews, interviewers obtained 30ml of whole blood that was collected in 4 X 7.5 ml (EDTA containing) monovettes. The blood samples were labelled, gently mixed, and stored frozen upright in a -20°C freezer pending extraction. Interviewers collected saliva derived DNA from twelve family members who refused to give blood using the Oragene saliva collection system and extraction kits (DNA Genotek Inc., Ontario, Canada). This is a robust non-invasive method of DNA collection that preserves the DNA for several months prior to extraction with no significant loss of yield. All bar five participants provided a blood or saliva sample following the interview, and twenty-five participants, seventeen of whom were under the age appropriate for interview (younger than 5 years old) provided a blood or saliva sample without completing an interview. Sufficient information was attained from first-degree relatives regarding the psychiatric history of family members who were not interviewed but provided blood or saliva samples (were taken to be unaffected by the interviewers but were assigned "unknown" status in the analyses). Overall there were three hundred and twenty four samples available for DNA extraction.

4.2 DNA Extraction

Genomic DNA was isolated from whole blood and saliva usually within two days of sampling by laboratory technicians at the Federal University of São Paolo using standard procedures. DNA samples were sent to King's College London on dry ice via express international mail for genotyping.

4.3 Affymetrix GeneChip® Mapping 10K Array2.0

Genotyping was conducted using the Affymetrix 10K microarray, which works on the principle of allele-specific hybridisation between DNA samples and 25-mer oligonucleotide probes. Each SNP is interrogated in both the sense and anti-sense strands with 40 different 25-mer oligonucleotides represented in five probe quartets that are tiled onto the microarray. Each probe quartet includes a perfect match for allele A (PM_A), a perfect match for allele B (PM_B), a one base pair (bp) mismatch for allele A (MM_A) and a one base pair mismatch for allele B (MM_B) (Wang et al. 1998) (Figure 4.1).

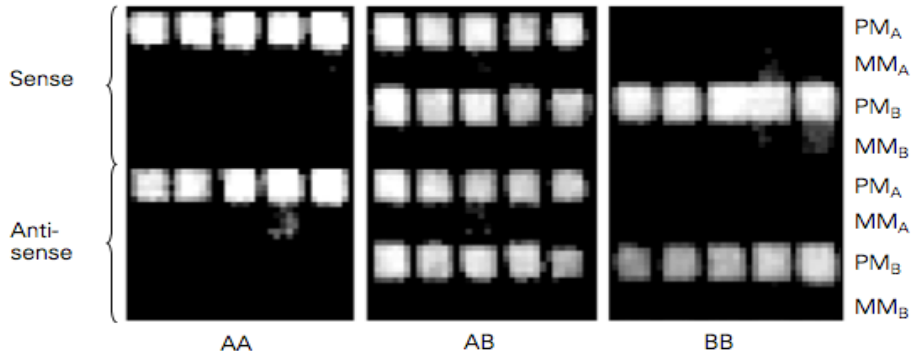


Figure 4.1. Scanned hybridisation signal intensities for a single SNP on the microarray. Three individuals are shown, one for each genotypic group (AA, AB and BB). Forty unique probes, twenty on the sense strand and twenty on the anti-sense strand represent each SNP on the microarray, and each probe is present in millions of copies. A hybridisation signal in only visible for the perfect match probes to allele A (PM_A) for the AA homozygote, and for the perfect match probe to allele B (PBM_B) for the BB homozygote. Figure from Affymetrix 10K mapping manual.

In this manner, the 10K SNP microarray allows the simultaneous interrogation of 11,555 SNPs distributed throughout the genome, with a mean inter-SNP distance of 210 kilobases (Kb) and average SNP heterozygosity of 0.37 (Kennedy et al., 2003). The SNPs on the 10K microarray were not selected on the basis of putative functionality or as informative 'tagging' SNPs. Rather computer predicted fragment

lengths based on Xba I enzyme restriction sites upstream and downstream of SNPs were used to select SNPs from the SNP Consortium repository in January and September 2001. This procedure identified 55,605 SNPs from which a subset of 11,555 was identified using selection criteria designed to ensure accuracy and reproducibility of the microarray assay (Holden, 2002).

4.3.1 The GeneChip® Mapping 10K 2.0 Assay Protocol

The BBF samples were processed according to the GeneChip® Mapping 10K 2.0 Assay Protocol, which was preformed over four days and followed eleven stages; DNA preparation, restriction enzyme digestion, ligation, PCR, purification and elution, quantification, fragmentation, labelling, hybridisation, washing staining and scanning, and finally data retrieval and genotype calling (Figure 4.2).

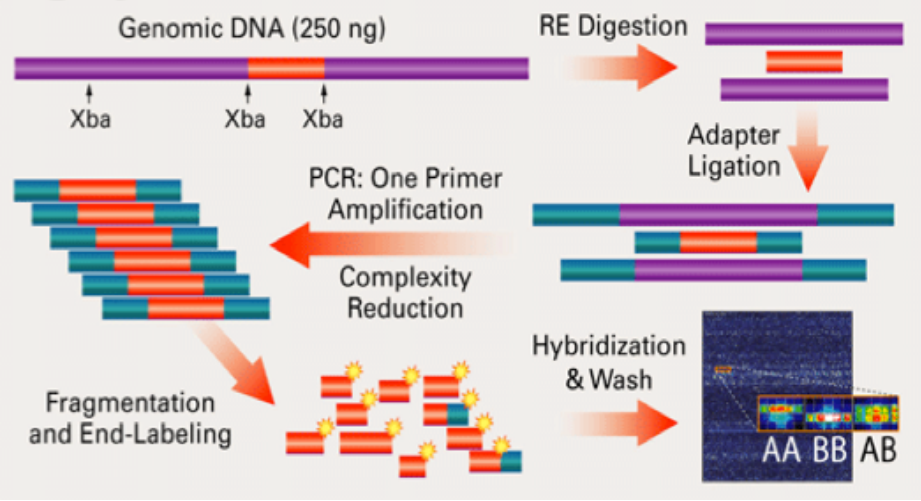


Figure 4.2 Overview of the GeneChip® Mapping 10K Assay protocol. Figure from (http://www.affymetrix .com)

Three hundred and twenty-four BBF samples were genotyped using this protocol. With every batch of samples genotyped, a reference DNA sample provided by Affymetrix was processed as a positive control to test the accuracy with which the protocol was completed and a sample of molecular biology grade water was

processed as a negative control in all the steps leading to labelling and hybridisation to assess the presence of contamination in the reactions. Three batches of 92 BBF samples were processed in this manner after initially processing three batches of 22 samples at a time. Some samples were run in duplicates (n=18). In our laboratory, we have 2 Affymetrix® GeneChip® Fluidics Stations (each containing 4 modules) which meant that a maximum of 8 microarrays could be washed and stained at a time. Therefore, three rounds of washing and staining were completed in one day after which the microarray were scanned together. The steps of the protocol are presented next.

Step 1: Genomic DNA Preparation

BBF DNA samples were quantified using a NanoDrop fluorospectrometer, which measures the amount of ultraviolet (UV) light that is absorbed as it passes through an aqueous solution. Concentrations of the DNA solution can be estimated by comparing the proportion of light passing through a blank (I_0) (water), relative to that which passes through the solution in question (I). From this two statistics are calculated, transmittance (T) and absorbance (A):

$$T=I/I_0 \qquad A= - Log_{10} \mathrm{T}$$

The absorbance statistic enables the concentration of DNA to be estimated when UV light is passed through at 260 nm; the wavelength absorbed by nucleic acid. To assess DNA purity, the intensity of absorbance of the DNA solution at 280nm; the wavelength absorbed by protein is estimated and the ratio 260:280 is used as a purity measure. A DNA sample with a 260:280 ratio of 1.8 is considered to be pure and free from protein contamination and a DNA sample that has a 260:280 ratio lower than 1.8 is said to be contaminated with proteins. The majority of BBF samples (n=280) were free from contaminants, however, a subset of samples scored below 1.4 indicating

protein contamination, most probably heme. These samples were cleaned using a standard phenol-chloroform and proteinase K DNA purification procedure. Following, two rounds of phenol-chloroform and proteinase K cleanup only 24 samples were usable. In total 304 samples were carried through to step 2 of the protocol.

Step 2: Restriction Enzyme Digestion

Genomic DNA (250ng at 50ng/µl) was digested with 10 units of the restriction enzyme Xba I in the following reaction; 10.5µl H_2O, 2µl 10X NE buffer, 2µl BSA at 1mg/ml, 0.5 XbaI at 20U/µl, for 2 hours at 37°C in the thermal cycler, followed by enzyme inactivation for 20 minutes at 70°C. This step of genomic fractionation is necessary to reduce the chances of cross-hybridisation and therefore non-specific allele signaling in subsequent steps.

Step 3: Ligation

The digested DNA fragments were ligated to Xba adaptors (Affymetrix, Santa Clara, CA) that recognise the cohesive four base pair (bp) overhangs, using T4 DNA Ligase in the following reaction; 1.25ul Adapter Xba I at 5uM, 2.5ul X10 T4 DNA Ligase buffer, 1.25ul T4 DNA Ligase at 400,000U/ml and 20ul digested DNA, for 2 hours at 16°C in the thermal cycler, followed by an enzyme inactivation step for 20 minutes at 70°C. All DNA fragments resulting from restriction enzyme digestion, regardless of size, are substrates for adaptor ligation. Ligated DNA was diluted 4X with water (25µl Ligated DNA and 75µl of water).

Step 4: Polymerase Chain Reaction (PCR)

A generic primer provided by Affymetrix that recognises the Xba I adapter sequence was used to amplify the adapter-ligated DNA fragments in a PCR reaction. Fragments

in the size rage 250-1000 bp are preferentially amplified; this narrow size range of amplicons reduces the complexity of the genome and allows allele-specific hybridisation of the samples to the probes on the microarray.

In this step, 90µl of PCR master mix containing 10µl PCR buffer, 10µl dNTPs (2.5mM each), 10µl $MgCL_2$ (25mM), 7.5µl PCR Primer Xba (10µM), 7.5µl AmpliTaq Gold (5U/µl) and 50.5µl H_2O,was mixed with 10µl of DNA. Each sample was amplified independently in four separate reactions to produce sufficient product for hybridisation to one microarray. The PCR reaction entailed cycling three steps: denaturation, primer annealing, and DNA synthesis, under the following conditions:

Temperature	Time	Cycles
95°C	3 minutes	1
95°C 59°C 72°C	20 seconds 15 seconds 15 seconds	35
72°C	5 minutes	1

Following the completion of the PCR, 3µl of each PCR product, mixed with 3µl of Gel loading dye, were run on 2% Agarose gel (prepared by adding 10g Agarose to 500mL TBE heated in the microwave until the Agarose was completely dissolved and the solution was clear and mixed with12µl of Ethidium Bromide) at 120 Volts for 1 hour to check that sufficient amplification of each sample in the correct size range had occurred. The samples were stored at -20 until the next day.

Step 5: PCR Purification and Elution with QIAGEN MinElute 96 UF PCR Purification Plate

The PCR products were purified to remove excess primers, unincorporated dNTPS. The four PCR reactions for each sample were pooled into one well of the MinElute plate (QIAGENInc., Hilden, Germany) (total 400µl amplified DNA product per well)

and dried down using a vacuum, which was maintained at approximately 800 mbar for five hours. Once the wells were dry, the PCR products were washed by adding 50µl of molecular biology grade water, and dried completely using the vacuum for approximately two hours. This step was repeated twice for a total of three water washes, after which the MinElute plate was removed from the vacuum and tapped on a stack of absorbent paper to remove any liquid remaining on the bottom of the plate. The resulting DNA products were reconstituted in 40µl of EB buffer and shaken on a plate shaker at 1000 rpm for 5 minutes. Purifying the PCR products took approximately 7 hours to complete as opposed to the expected 1 hour 50 minutes suggested by Affymetrix. I suspect that was the case because a suboptimum vacuum was used to complete this step. The samples were stored at -20 for completion of the protocol the next day.

Step 6: Quantification of Purified PCR Product

A UV spectrophotometer was used to quantify the purified PCR products. A forty-fold dilution (4µl purified PCR product diluted in 156µl of molecular grade water) was prepared in an optical plate and read at an absorbance of 260nm. The PCR products were quantified in triplicate and a water blank was included as a measure of accuracy. The purified PCR products were normalised to concentrations of 20µg per 45µl solution by adding more EB buffer and then transferred to a new plate in preparation for fragmentation. In cases where there were less than 45µl of DNA available, EB buffer was added to a final volume of 45µl, irrespective of concentration.

Step 7: Fragmentation

To allow sufficient hybridisation of the DNA molecules to the microarray probes, the purified PCR products were fragmented to approximately 50bp to 100bp using Deoxyribonuclease I (Affymetrix), an endonuclease, which was diluted and added to 45µl of purified PCR product and 5µl of fragmentation buffer (Affymetrix). The samples were fragmented in a pre-heated thermal cycler at 37°C for 30 minutes, followed by 95°C for 15 minutes. Following the completion of the program, 4µl of each fragmented PCR product mixed with 4µl gel loading dye, was run on a 4% pre-cast gel at 120 Volts for one hour to ensure fragmentation occurred successfully. The size of fragments was on average less than 180 bp (Figure 4.3).

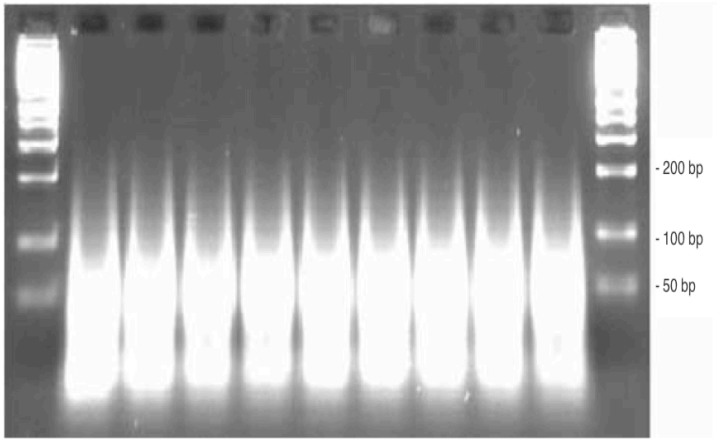


Figure 4.3. A typical example of fragmented PCR products from the BBF run on 4% gel at 120 Volts for one hour. The average fragment size is less than 180 bp.

Step 8: Labelling

The fragmented DNA products were end-labelled with biotin using Terminal Deoxynucleotide Transferase (TdT, Affymetrix) to allow attachment of florescent molecules during the staining procedures that follow. The biotin was prepared in a master mix containing 14µl TdT buffer (5X), 2µl GeneChip® labelling reagent at (5mM) and 3.4µl TdT. Aliquots of 19.4µl master mix were added to 50.6µl of

fragmented DNA and the reaction was heated at 37°C for 2 hours, followed by 95°C for 15 minutes.

Step 9: Target Hybridisation

The microarrays were taken out of the cold room where it was stored at 4°C, and left on the bench for approximately 30 minute to equilibrate to room temperature. This was done to prevent the microarray rubber septa from cracking, a major cause of unwanted leaking during hybridisation. Biotin labelled DNA was added to 190µl hybridisation mix containing 12µl 12X MES stock, 13µl DMSO, 3µl X50 Denhardt's solution, 3µl EDTA (0.5M), 3µl Herring Sperm DNA (10mg/ml), 2µl X1 oligonucleotide control reagent, 3µl Human Cot-1 (1 mg/ml), 1µlTween-20 (3%), and 140µl Tetramethyl Ammonium Chloride (5M). This mixture was heated at 95°C for 10 minutes to denature, cooled quickly on ice, and then briefly spun. It was then placed in a hybridisation oven at 48°C for 2 minutes before being injected into the microarrays and hybridised overnight for 18 hours at 48°C. Loading the microarray was done by inserting a pipette tip into the upper septum of the microarray and injecting the hybridised DNA into the bottom septum of the microarray. We injected 120µl of the hybridisation mix into the microarray, more than the 80µl recommended by Affymetrix, to prevent the formation of bubbles that interfere with the hybridisation process. The microarray septa were then sealed with tough-spots (Figure 4.4).

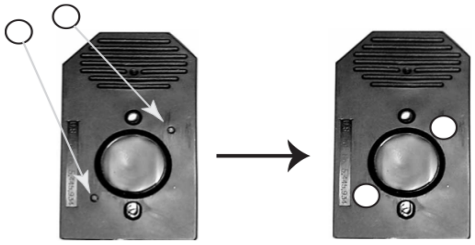


Figure 4.4. Affymetrix 10K Microarray. The arrows point to septa used to fill the microarray.

Step 10 Washing, Staining and Scanning Arrays

Following 18 hours of hybridisation, the hybridisation mix was extracted from the microarray using a pipette. A stringent wash protocol was then performed to remove any remaining non-hybridised DNA fragments from the microarray that would otherwise cause background noise during scanning. The microarrays were washed and stained in a three-stage process consisting of two washes using a non-stringent wash A (6X SSPE, 0.01% Tween 20) and a stringent wash B (0.6X SSPE and 0.01% Tween-20), a streptavidin phycoerythin (SAPE) stain containing 450µl stain buffer (6X SSPE, 0.01% Tween-20, 1X Denhardt's solution) and 5µl SAPE(1mg/ml),an antibody stain with biotinylated antibody containing 450µl stain buffer(6X SSPE, 0.01% Tween-20, 1X Denhardt's solution)and 5µl biotinylated antibody (0.5mg/ml) followed by a final stain with SAPE.

Using Fluidics station_450 and Affymetrix GeneChip® Operating Software the arrays were washed and stained using the following protocol:

Post hybridization wash 1: 6 cycles of 5 mixes/cycle with wash buffer A at 25^{o}C
Post hybridization wash2: 6 cycles of 5 mixes/cycle with wash buffer B at 45^{o}C
Stain 1: Stain the array for 10 minutes in SAPE solution at 25^{o}C
Post stain wash: 6 cycles of 5 mixes/cycle with wash buffer A at 25^{o}C
Stain 2: Stain the array for 10 minutes in antibody stain solution at 25^{o}C
Stain 3: Stain the array for 10 minutes in SAPE solution at 25^{o}C
Final Wash: 10 cycles of 6 mixes/cycle with wash buffer A at 30^{o}C.

Following the completion of the wash and stain cycles, the microarrays were filled with a 1X array holding buffer containing 8.3ml MES buffer (12X), 18.5ml NaCL (5M), 0.1ml Tween-20 and 73.1ml water) in preparation for the scanner. Any excess fluid was cleaned from around the septa, and tough spots were reapplied to each

septum. The arrays were placed in the GeneChip® Scanner 3000 7G and the GCOS software was used to the scan the arrays.

Step 11: Data Generation and Genotype Calling

The scanner automatically generated .CEL files, containing raw allele intensity scores, which were processed by the GeneChip® DNA Analysis Software (GDAS) to derive SNP genotypes. A Modified Partitioning Around Medoids (MPAM) mapping algorithm that assigns SNP genotypes on the basis of relative allele signal (RAS) intensities is used by the software (Liu et al., 2003). The SNP allele signals are scored relative to a centre point (medioid) given to each AA, AB, and BB genotype. The mediods are based on samples from multiple individuals from diverse populations and reflect the allele frequencies of the SNPs (Affymetrix manual). To calculate the RAS score for each probe quartet, the average mismatch (MM_{Ave}) from the expected mediod intensity is first calculated using:

$$MM_{Ave} = (MM_A + MM_B)/2$$

Where MM_A refers to the mismatch intensity value for allele A and MM_B refers to allele B. MM_{Ave} is then subtracted from the perfect match intensity for allele A (PM_A) and the perfect match intensity for allele B (PM_B) to correct for background noise due to non-specific hybridisation, as follows:

$$A = \max (PM_A - MM_{Ave})$$

$$B = \max (PM_B - MM_{Ave})$$

Finally the RAS score is calculated using

$$RAS_{quartet} = A/A+B$$

These calculations are repeated for the sense and antisense strands for each quartet and a median RAS score for the sense strand (RAS_1), a median RAS for the antisense strand (RAS_2) is derived. RAS_1 and RAS_2 are plotted for each SNP using MPAM. If (RAS_1,RAS_2) cluster near (0,0) genotype BB is called, if they cluster near (1,1) genotype AA is called, and if they cluster near (0.5,0.5) genotype AB is called. RAS scores falling outside the clustering boundaries are not assigned genotypes (Figure 4.5).

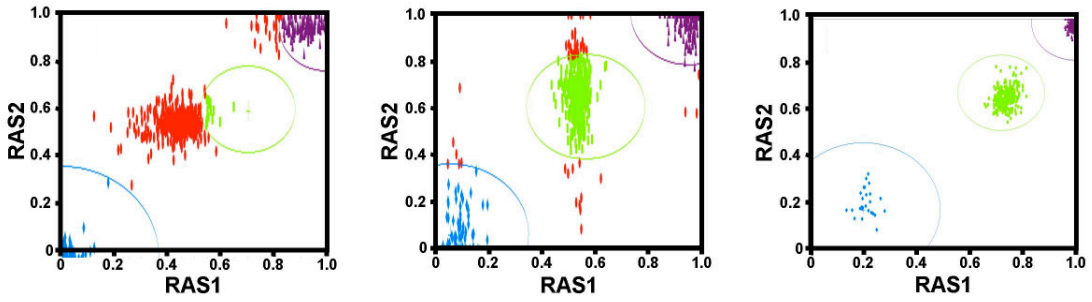


Figure 4.5. A graphical representation of three SNPs. The blue ellipse represents the area of BB genotypes near (0,0), the green ellipse represents the area of AB genotypes near (0.5,0.5) and the purple ellipse represent the area of AA genotypes near (1,1). SNP with relative allele scores falling within those areas are assigned the appropriate genotypes. SNPs with relative allele scores falling outside those areas, represented here in red, are not assigned a genotype call. Each SNP has a different area for AA, AB, and BB genotypes reflective of different population allele frequencies. Figure adapted from (Huentelman et al., 2005).

4.4 Quality Control (QC)

The quality of genotype data is a major source of bias and loss of power in linkage and association studies making stringent quality control (QC) an essential step prior to any analysis of genetic variation. It is standard in family based studies to conduct per-individual QC and per-SNP QC, in that order, as well as Mendelian and non-Mendelian error checks.

4.4.1 Genotyping Quality Control

A considerable number of genotyping errors occur in most large genotype datasets. This mistyping could be due to human oversight, to shortcomings in genotype scoring software, or simply to errors in the biochemical assays. Several authors have shown that genotyping errors mask linkage signals and distort the accuracy of the marker map by giving an inaccurate estimate of the disease location on the marker map. In two-point linkage analysis genotyping errors usually lead to inflated estimates of recombination fractions, whereas in multipoint linkage analysis they result in the false exclusion of true linkage to disease (Göring & Terwilliger, 2000). Even a small (1% to 2%) error rate can have an enormous impact on linkage results (Abecasis, Cherny, & Cardon, 2001; Buetow, 1991; Sobel, Papp, & Lange, 2002). It is therefore essential to obtain a data set as free of genotyping errors as possible before proceeding to linkage analysis.

4.4.1.1 Per Individual Quality Control

Data quality control was commenced by excluding individuals with more than 10% missing genotype rate. The lowest call rates (n=12) were obtained from individuals in Branches 2 and 3 of the family, where the quality of the DNA was not optimal and

several DNA cleaning steps were performed *(section 4.3.1)*. These samples were collected and batch processed, so it is suspected that handling (repeat freeze thawing) as well as variation in laboratory techniques may have lead to impediments to optimal assay performance. In addition, in the first wave of ascertainment, when Branch 1 of the BBF was collected, psychiatrists visited Senhora de Oliveira or nearby cities for a maximum of two days before returning to São Paulo and processing the DNA samples with laboratory technicians. However, in the second wave of ascertainment, when Branches 2 and 3 were collected psychiatrists spent more days in the village and thus delayed processing the DNA samples by up to a week.

After the exclusion of low quality individual data, the mean call rate in the BBF samples was 93%. Discussion with Affymetrix customer support revealed that this array has a lower average call rate than the Affymetrix 500k and 1M arrays, being one of the earlier designed arrays in the Affymetrix product range.

Gender misspecification was then addressed by looking at the homozygosity of X-linked SNPs using PLINK v1.07 (Purcell et al., 2007) sex check, which calculates an inbreeding coefficient estimate (F Co-efficient), based on the observed versus expected number of homozygous X-chromosome SNPs. This highlights individuals for whom reported sex in the pedigree file does not match X-chromosome estimated sex. This test is conducted mainly to identify sample mix-ups and errors in pedigree data entry. Because males only have one copy of the X-chromosome they cannot be heterozygous for any marker outside the pseudo-autosomal region of the Y-chromosome. As a result, one expects male samples to have a homozygosity rate of 1 (PLINK makes a male call if F co-efficient is greater than 0.8 to account for genotyping error) and female samples to have a homozygosity rate of less than 0.2. Male samples that are mistakenly marked as female in the pedigree file will have a

higher than expected homozygosity rate and female samples marked as male will have a lower than expected heterozygosity rate (Anderson et al., 2010) (Table 4.1).

ID	Assigned Sex	Genotyped Sex	Status	F Co-efficient
54	2	2	Ok	0.11
1	1	1	Ok	1
682	2	1	Problem	1
34	2	0	Ambiguous	0.27
37	2	0	Ambiguous	0.34
39	2	0	Ambiguous	0.27
97	2	0	Ambiguous	0.55
98	2	0	Ambiguous	0.35
185	2	0	Ambiguous	0.48

Table 4.1. PLINK sex check used as a quality control step mainly to identify sample mix. The F Co-efficient for individuals with the correct sex, problematic sex and ambiguous sex are presented. Males are expected to have an F Co-efficient greater than 0.8 and females are expected to have an F Co-efficient less than 0.2.

Sample 682 had discordant sex information and was removed from the data set and further investigation revealed it must have been mislabelled as no record of it existed. There were a number of ambiguous sex estimations, which belonged to females from consanguineous marriages. These samples were not removed from the dataset as comparison with their autosome inbreeding statistics revealed these females to be of high inbreeding/homozygosity, thus artificially inflating the X chromosome F statistic. After per individual quality control a total of 309 family members were left in the pedigree dataset.

4.4.1.2 Per SNP Quality Control

Following per-individual quality control, we commenced per-SNP quality control by excluding SNPs with more than 10% missing genotype rate (n=1133) using PLINK. We then identified SNPs with a minor allele frequency less than 25% in unaffected BBF founders (n=54), i.e. unrelated individuals who marry into the family or families under study, and excluded them from BBF pedigree dataset. This was done to increase the heterozygosity of the data and therefore information content.

4.4.1.2.1 Hardy Weinberg Equilibrium

In performing genetic studies it is normally assumed that unaffected individuals (or in the case of a population study the sample as a whole) are in Hardy-Weinberg Equilibrium (HWE). HWE states that under conditions of random mating, absence of migration, mutation, and selection at the gene in question, a biallelic marker with alleles A and a should have a stable frequency distribution of genotypes AA, Aa, and aa in the proportions p^2,$2pq$, and q^2 respectively where p is the frequency of allele A and q is the frequency of allele a. HWE may therefore be calculated for a biallelic marker using the following equation:

$$p^2 + 2pq + q^2 = 1$$

The most common cause of significant deviations from HWE is genotyping errors in the laboratory or due to genotype calling algorithms that consistently mis-score certain heterozygotes as homozygotes or fail to score a specific allele. Testing for HWE is therefore part of quality control procedures in gene mapping studies and is typically performed using a standard chi-square test that compares the expected allele frequencies with those observed in the data (significant p-values would indicate deviation from HWE).

HWE is only expected to occur in randomly mating populations as both inbreeding and assortative mating cause increases in homozygosity (in all genes and genes involved in the trait that is assortatively mated respectively) and therefore deviations from HWE. As both inbreeding and assortative mating are evident in the BBF only unaffected BBF founders were used to delineate SNPs that deviated from HWE, which were subsequently removed from the entire pedigree data. Both of these

procedures were conducted using PLINK. After per SNP quality control 5315 SNPs were available for analysis.

Following the conclusion of standard per-individual and per-SNP QC checks for pedigree structure, Mendelian and non-Mendelian errors were conducted.

4.4.2 Pedigree Structure Errors

Pedigree structure errors occur due to a number of factors including unreported adoptions, non-paternity, errors in entry of pedigree information, and sample mix-ups. We used PLINK pair-wise Identical-By-Decent (IBD) estimation across all SNPs to detect misspecified family relationships. The expectation is that IBD = 1 is for duplicates or monozygotic twins, IBD = 0.50 for first-degree relatives, i.e. parents, children, siblings, IBD=0.25 for second-degree relatives, i.e. half-siblings, grandparents, uncles/aunts, cousins, and IBD=0.125 for third-degree relatives, i.e. second cousins, great uncles/aunts. Evaluation of pair-wise IBD estimates between BBF members revealed a number of misspecified family relationships, which were dealt with on an individual basis.

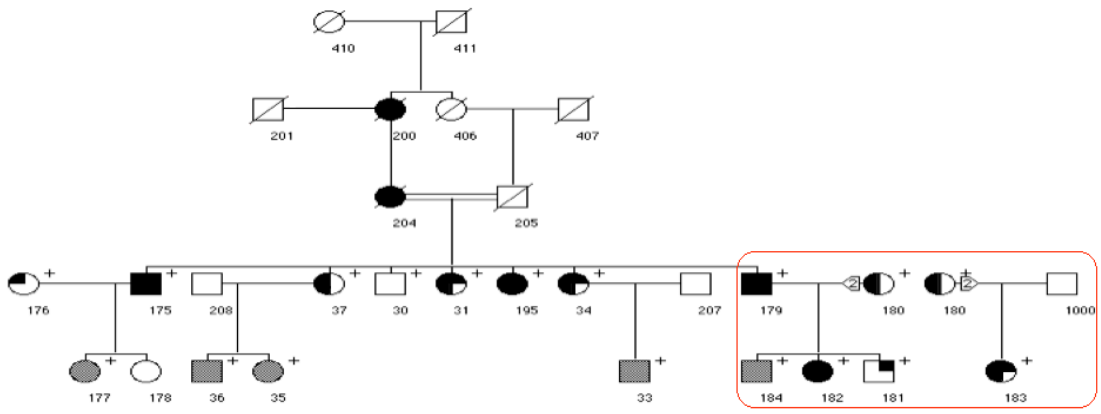


Figure 4.6. Subfamily 3 with relationship misspecifications in the section highlighted in red, as determined by IBD estimates.

In subfamily 3 (Figure 4.6), non-paternity was established for individual 183 based on estimated allele sharing of zero between her and her alleged father (Table 4.2). As she had half siblings, a dummy paternal ID (1000) was created and her data was included in all subsequent analyses.

ID 1	ID2	Relationship	Pair-wise IBD	Conclusion
183	179	Father	Zero	Non-paternity
183	180	Mother	0.5	Mother
183	181	Sibling	0.28	Half-sibling
183	182	Sibling	0.22	Half-sibling
183	184	Sibling	0.23	Half-sibling

Table 4.2. IBD estimates for a section of subfamily 3 showing family relationship misspecification

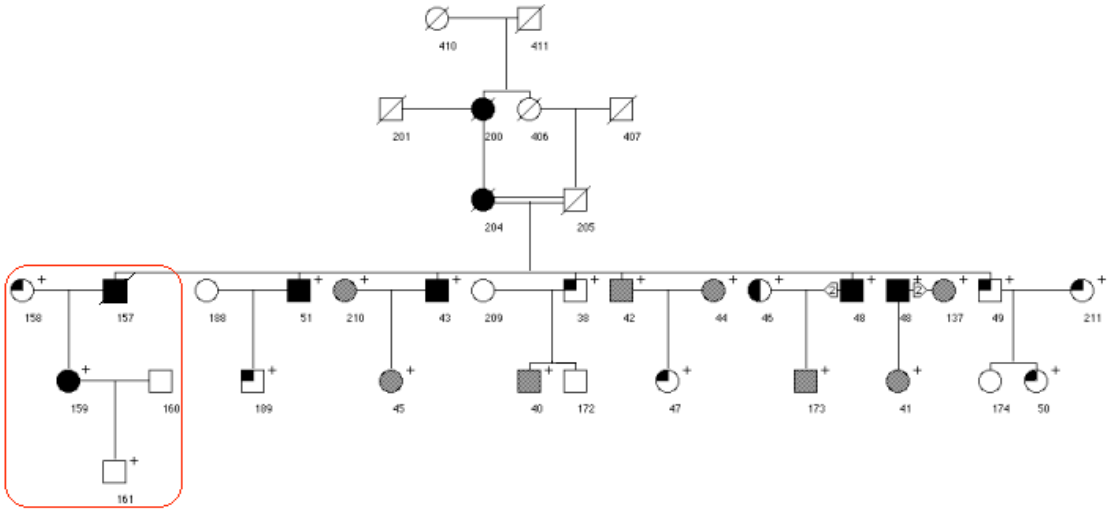


Figure 4.7. Subfamily 2 with relationship misspecifications in the section highlighted in red, as determined by IBD estimates.

A similar picture emerged from subfamily 2, where the section of the family, highlighted in red in Figure 4.7 was removed from the pedigree data due to non-paternity. As the father in question was deceased, we evaluated the relationship of his daughter 159 with supposed uncles (Table 4.3), and found that the genetic data indicated no relationship to her. In fact individual 159 had no genetic relationship to

any BBF member. As a result, individuals 158, 159, 160, and 161, were removed from the data set.

ID 1	ID2	Relationship	Pair-wise IBD	Conclusion
159	158	Mother	0.5	Mother
159	161	Son	0.5	Son
159	43,38, 42,48,49,51	Uncles	0	No relationship

Table 4.3. IBD estimates for a section of subfamily 2 showing family relationship misspecification determined by IBD estimates.

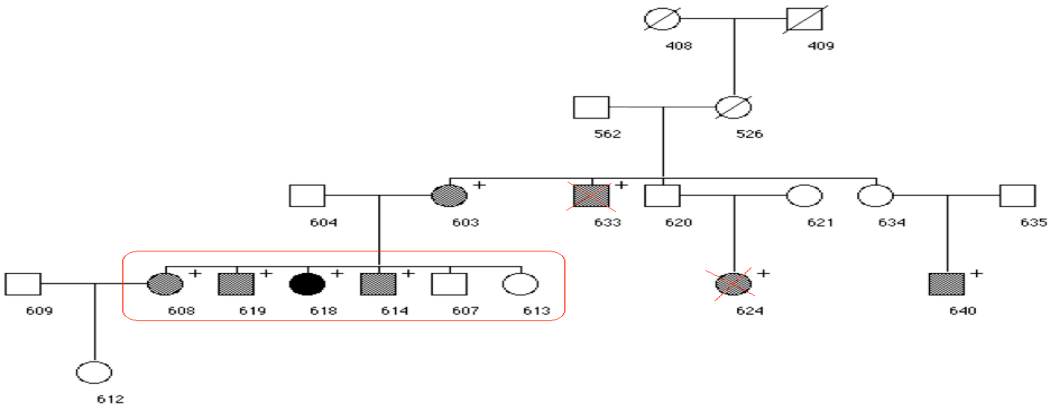


Figure 4.8. Subfamily 17 with relationship misspecifications in the section highlighted in red, as determined by IBD estimates. Individuals 633 and 624 had low genotyping rates.

Contrary to the first two discrepancies in family relationships, family relationships in subfamily 17 (Figure 4.8) were more difficult to decipher and the whole subfamily was dropped from all subsequent analyses. The decision to do so was backed by inconclusive relationships between the children of 603, who in the diagnostic interview discussed having multiple extra-marital affairs (Table 4.4).

ID 1	ID2	Relationship	Pair-wise IBD	Conclusion
608	614	Siblings	0.5	Siblings
608	618	Siblings	0.14	Ambiguous
608	619	Siblings	0.124	Ambiguous
614	618	Sibling	0.15	Ambiguous
614	619	Sibling	0.25	Ambiguous
618	619	Sibling	Zero	Ambiguous
633	Low Genotyping Rate			
640	Low Genotyping Rate			

Table 4.4. IBD estimates for a section of subfamily 17 showing family relationship misspecification

4.4.3 Mendelian Inconsistent Error

When dealing with family data an important part of the genotyping quality control is testing for Mendelian inconsistencies, or genotype patterns that do not follow Mendel's law of inheritance, which states that every individual inherits one paternal allele and one maternal allele. Mendelian inconsistencies help identify genotyping errors or pedigree structure errors. After the correction of pedigree structure errors informed by PLINK pair-wise IBD estimates, the program PEDSTATS (Wigginton & Abecasis, 2005) was used to identify genotypes that did not conform to Mendelian inheritance in the pedigree, which were set to zero in the analyses.

4.4.4 Non-Mendelian Errors

Not all genotyping errors cause Mendelian inconsistencies. A non-Mendelian error is a genotyping error that is consistent with Mendelian inheritance. Simulation studies have shown the percentage of genotyping errors consistent with Mendelian inheritance could be up to approximately 40% for multiallelic markers and as high as 87% for biallelic markers (Douglas, Skol, & Boehnke, 2002). As a result a check for Mendelian inconsistencies alone is unlikely to identify all problematic genotypes. Non-Mendelian errors are much harder to determine than Mendelian errors, however, they still have a profound effect on the validity of linkage analyses.

Non-Mendelian genotyping errors are detected when there is excessive number of recombination events occurring in a particular chromosomal location. Two or more close recombination events on the same chromosome are uncommon, due to interference, where the occurrence of a crossover event alters the probability of another crossover event occurring at a nearby location. This probability could be increased or decreased but the latter is more usual as studies have suggested the probability of double recombinants within a 20 centiMorgan (cM) interval in humans is only 2 in 1000 (Broman & Weber, 2000). Therefore, the occurrence of a double recombinant more likely reflects genotyping error in one of the markers used to infer the location of the recombination event.

The error detection procedure in the program MERLIN (Abecasis, Cherny, Cookson, & Cardon, 2002) was used to search for non-Mendelian errors or unlikely recombinants in the dataset. MERLIN considers all pedigree data simultaneously, i.e. not only pairs of individuals, so its error detection procedure has improved accuracy in larger pedigrees. Unlikely genotypes were flagged and MERLIN "pedwipe" was used to set them to zero in the pedigree dataset. McLinkage (Thomas, Camp, Farnham, Allen-Brady, & Cannon-Albright, 2008) "CheckErrors" procedure was also used to calculate the posterior probability of genotype mistyping at each observed genotype in the data, which were incorporated into subsequent McLinkage analyses to account for genotyping errors.

4.5 Data Analysis and Statistical Methods

4.5.1 Splitting the BBF Sub-families for Analysis

Algorithms for exact LOD score calculations (e.g. Elston-Steward, Lander-Green) consider each inheritance configuration consistent with the available genotype data to calculate LOD scores. These algorithms are limited by the complexity and size of the pedigrees they could handle because the total number of underlying inheritance configurations becomes too large to compute. A solution is to split large pedigrees into computable subfamilies that meet the complexity constraints of exact linkage analyses. Linkage programs, such as MERLIN, GENEHUNTER (Daly et al., 1998), or Mendel (Lange et al., 2001) that use the Lander-Green exact LOD score algorithm require large pedigrees to be split into subfamilies of no larger than 30 bits, where a bit is defined as two times the number of individuals with parents presented in the pedigree minus the number of pedigree founders. As the computational complexity of the Lander-Green algorithm increases linearly with the number of markers, but exponentially with the bit-size of the pedigree, it can analyse a large number of markers if the pedigree bit size is small. While splitting large pedigrees apart can reduce inheritance information by removing transmission links among pedigree branches, it is the only way to perform exact multipoint analysis in large complex pedigrees such as the BBF.

To accommodate the requirements of MERLIN, our linkage software of choice for exact LOD score calculations, the BBF was split into 19 appropriately sized (30 bits or less) subfamilies. When splitting the BBF important aspects of family relatedness, such as first cousin marriages and complex marriage loops were

preserved where possible and the number of duplicate individuals in the subfamilies was kept to a minimum *(See Chapter 2 section 2.17).*

4.5.2 Pedigree Datasets

Whole genome linkage analyses were separately conducted on two sets of data: Branch 1 and the BBF, which included Branches 1, 2, and 3. The decision to analyse Branch 1 separately and as part of the BBF was based on the fact that Branch 1 appeared to be a distinct group that manifested a higher density of psychiatric disorders. Both clinically and genotypically Branch 1 seemed to be more homogeneous than the other two branches of the BBF. Branches 2 and 3 were more outbreed than Branch 1 and segregated milder forms of BPD and mood disorders in general. However, analysing the BBF, it was thought would offer more power to detect linkage, as more recombination would be available for the analyses.

Following the completion of the described QC procedures, 309 individuals with 5315 genotyped SNPs were available for the linkage analyses. Four clean pedigree files were generated in standard linkage format:

1. Branch 1 subfamilies including 208 genotyped individuals and 61 un-genotyped individuals split into 12 subfamilies (n=269).
2. Branch 1, Branch 2, and Branch 3 subfamilies including 309 genotyped individuals and 91 un-genotyped individuals split into 19 subfamilies (n=400), referred to hereinafter as the BBF subfamilies.
3. Branch 1 with structure in tact including 208 genotyped individuals and 34 un-genotyped individuals (n=242)
4. BBF with structure in tact including 309 genotyped individuals and 52 un-genotyped individuals (n=361).

Parametric and non-parametric linkage analyses were performed on the four pedigree files using four phenotype models:

1. Narrow - included family members diagnosed with BPI, BPII, or SAD.
2. Broad – included family members in the narrow model plus those diagnosed with BP NOS and cyclothymia.
3. Super - included family members in the broad model plus those diagnosed with one episode of major depressive disorder, recurrent episodes of major depressive disorder, and cyclothymia.
4. Depression – included family members diagnosed with unipolar depression and cyclothymia only.

Details of the parametric and nonparametric methods will be presented next.

4.5.3 Linkage Information content

Information content mapping measures the total inheritance information extracted by the map of markers used in a study. It is a function of marker heterozygosity and the number of detectible meioses (which allow identification of whether a gamete is recombinant or non-recombinant) in the pedigree(s) under study. For multipoint linkage analysis, information content is also a function of marker density and spacing (Kruglyak, 1997). The information content is essentially how much power you have to assess any evidence for linkage. High information content is essential for genome wide searches for disease susceptibility loci or other traits so that regions of no linkage can be excluded, regions of significant linkage can be detected, and the linkage interval can be accurately defined (Kruglyak, 1997).

MERLIN was used to calculate the inheritance information captured by the SNP data set in Branch 1 and the BBF using an entropy based measure (Kruglayk, 1996). According to this measure information content of one indicates the availability

of full information on the inheritance pattern in a pedigree(s), where as information content of zero indicates total uncertainty about the inheritance pattern.

4.5.4 Parametric Linkage Analysis

The LOD score is a standard statistic used for linkage analysis. It compares the likelihood of the data under linkage ($\theta < 0.5$) with the likelihood of the data under no linkage ($\theta = 0.5$)*(see chapter 3 section 3.2).* To analyse the data under the LOD score method an inheritance model for BPD representative of the population under study was specified, SNP marker allele frequencies were estimated and recombination fractions between SNP markers were adjusted as will be discussed in this section.

4.5.4.1 The BPD Inheritance Model

In parametric linkage analysis of complex diseases, such as BPD, where the mode of inheritance is unknown, it is common practice to analyse the data under multiple inheritance models, most typically dominant, recessive, or intermediate models with incomplete penetrance. Taking as a precedent the large family studies from the Central Valley of Costa Rica and Colombia (Herzberg et al., 2006; McInnes et al., 1996; Service et al., 2006), the BBF data was analysed under a recessive and a dominant inheritance model with incomplete penetrance. We assumed 1% penetrance for zero copies of the disease allele (i.e. phenocopy rate), 81% penetrance for one copy of the disease allele, and 90% penetrance for two copies of the disease allele, thus giving a penetrance vector of (0.01, 0.81, 0.90) for the dominant model and (0.01, 0.01, 0.90) for the recessive model.

To calculate the frequency of the disease allele, we used BPD population prevalence rates of 1.5%, 5%, and 20% for the narrow, broad, and super phenotype models respectively. These population prevalence rates are consistent with

epidemiological surveys from European populations (Kessler, Chiu, Demler, & Walters, 2005; Judd & Akiskal, 2003). To calculate the disease allele frequencies for the various phenotype models, we used the general two allele single major locus formula

$$K_p = f_1 p^2 + f_2 2pq + f_3 q^2$$

Where K_p is the population prevalence of the disorder, p is the frequency of the common "normal" allele, q is the frequency of the disease allele, $p + q = 1$, and f_1, f_2, and f_3 are the penetrances for zero, one, and two disease alleles. Under a dominant mode of transmission, the parameters we specified resulted in a disease allele frequency (q) of 0.003, 0.03, and 0.13 for the narrow, broad, and super phenotype models respectively. Under a recessive mode of transmission, the values of q were 0.07, 0.03, and 0.46 for the narrow, broad, and super phenotype models respectively.

4.5.4.2 The Depression Inheritance Model

Analyses on a depression only model were performed. The following parameters corresponding to a population prevalence for depression of 5%, and penetrance estimates of 50% for one or two copies of the disease allele, we specified disease allele frequency of 0.005 and parameters (0.05, 0.50, 0.50) for a dominant disease model and disease allele frequency of 0.033 and parameters (0.05, 0.05, 0.5) for a recessive disease model. These disease models have been previously used for the analysis of depression in multiple nuclear pedigrees from the UK (McGuffin et al., 2005).

4.5.4.3 The Marker Allele Frequencies

Accurate marker allele frequency estimation is necessary for the elimination of both false-positive evidence for linkage and biases in recombination fraction estimates. This is of particular importance when the number of marker alleles is large (Freimer, Sandkuijl, & Blower, 1993). Erroneous estimations are particularly problematic when genotypes for unavailable family members, particularly founders, are reconstructed in a dataset: a very common step in most linkage analyses. When affected offspring share a particular marker allele, the assumption is they either inherited it from a common ancestor, so it is identical-by-descent, or inherit it from multiple ancestors, and it is identical-by-state. Only the true population frequency of the shared marker allele could portray the true inheritance. When the frequency of the marker allele is underestimated, the probability of identity-by-descent, under linkage, will be overestimated (Ott, 1999). To estimate marker allele frequencies for the BBF study, we had three options:

1. To genotype a number of unrelated individuals from the population under study and to use their marker allele frequencies, an option that usually involves additional genotyping costs, and is therefore unsatisfactory.
2. To use the genotypes from the whole family, which typically leads to inaccurate estimation of population marker allele frequencies given that the genotypes in a family are not independent from each other.
3. To use the genotypes of BBF founders to estimate marker allele frequencies.

We opted for the third option. For analyses conducted using MERLIN estimates were derived by the program by counting marker allele frequencies among the BBF founders only using the --ff flag. For analyses conducted using McLinkage,

GeneCountAlleles, a program available as part of the McLinkage package was used. It is based on the C A B Smith's gene counting method, and works by iteratively inferring the genotypes of the founders in a pedigree given the current allele frequency estimates, the pedigree structure, and any genotypes observed in the pedigree and uses the posterior distribution of the founder allele frequencies to derive new maximum likelihood estimates for the marker alleles (Thomas & Camp, 2006).

4.5.4.4 The Recombination Fractions between Markers

The Affymetrix 10K annotation file contains the genetic positions, also known as cM positions of the 10K markers. These positions required adjustment for the following reasons. The genetic map measures the distance between genes for which one meiosis in 100 is recombinant. The basic assumption behind this measure is not entirely correct given that the probability of recombination is not uniform along the entire length of the chromosome. Recombination is inhibited in some regions, e.g. near the centromere, and increased in other regions, referred to as recombination 'hot spots.' In addition, positive interference, which we previously described, whereby the occurrence of a crossover reduces the likelihood of another crossover in its vicinity is a well documented phenomena that changes the proportion of crossovers in a chromosome (negative interference, whereby the occurrence of a crossover enhances the occurrence of other crossovers in the same region of the chromosome, e.g. double crossover, occurs with less likelihood than positive interference and also has an effect). Moreover, the number of chiasma in a bivalent segment, or two homologous chromosomes paired during the prophase of the first meiotic division, could interfere with the recombination frequency (Cavalli-Sforza, 2001). As such, the Kosambi mapping function which adjusts the genetic map distance based on interference, was used to re-calculate marker cM position, as follows:

$$M = 1/2 \tanh (2r)$$

Where r is the recombination fraction and M is the distance between markers in Morgans, which can be converted into cM by multiplying with 100. This step was conducted to ensure the use of a genetic map with relatively accurate inter-marker distances.

The 10K microarray SNP annotation were based on build 18 of the human genome adjustments to SNP physical positions using build 19 of the human genome were made and SNPs that did not map to the newer build were removed.

4.5.4.5 Analyses using MERLIN

Multipoint linkage analysis was conducted on 18 subfamilies from the BBF and 12 subfamilies from Branch 1 of the family under a recessive and dominant mode of disease transmission using the narrow, broad, super, and depression phenotype models. Heterogeneity LOD (HLOD) scores as well as LOD scores were calculated. As previously mentioned, linkage is affected by locus heterogeneity so the presence of more than one susceptibility locus in different families under study affects the power to detect linkage and HLOD scores allow for the detection of linkage even when a subset of the data shows no linkage between the disorder and marker(s) in question. MERLIN parametric linkage analyses report LOD scores, and HLOD scores with associated alpha (α) parameters, where an alpha of one indicates that all family members show evidence for linkage at a particular locus and an alpha of zero indicates that no family members show evidence for linkage. The per family option (using the –perFamily flag) was used to get LOD scores for each subfamily as well as total LOD scores for all subfamilies. This option was used to give a detailed account

of each subfamily's contribution to the overall LOD score and to evaluate if one or more subfamilies were driving the majority of the linkage signals detected.

4.5.4.6 Analyses using McLinkage

Multipoint linkage analyses of Branch 1 and the BBF were conducted using McLinkage. The program is based on the Markov Chain Monte Carlo (MCMC) algorithm capable of estimating LOD scores in large pedigrees. It does so by considering the underlying inheritance configurations (based on the observed genotypes) in proportion to their likelihood. Thus configurations that are theoretically possible but highly unlikely (probably due to the large number of recombinations they would require) will often not be considered. In the MCMC sampling from inheritance configurations begins randomly at any 'legal' or probable configuration and continues in a sequence of random sampling until a desired inheritance distribution that approximates the actual distribution is reached. The main difficulty with linkage programs based on the MCMC is deciding the number of samples needed to converge to a stationary distribution (Lange & Sobel, 1991).

By utilising the MCMC algorithm, McLinkage calculates multipoint LOD scores (MLOD) and Theta LOD (TLOD) scores between the specified disease phenotype and genetic markers in the pedigree. In a TLOD score, the inheritance configurations in the pedigree are determined based on all the genetic markers but the LOD score is estimated at each marker position at recombination fraction (θ) values of 0.0, 0.01, 0.1, 0.2, 0.3, 0.4, and 0.5 thus emulating a two-point linkage approach (that allows maximising the LOD score over the recombination fractions)(Abkevich et al., 2001).

The McLinkage analyses were conducted per chromosome using an error rate of 3%, which was determined using CheckErrors, a program available as part of the McLinkage package that calculates the posterior probability of errors in the data based on the observed genotypes. A 1000 random samples and 100 burn-ins (iterations at the beginning of the MCMC run that are discarded from the analysis with the aim of achieving a more stationary distribution) were initially employed. However, the reported LOD scores were inconsistent, indicating that the sampling distribution did not converge. In fact each analysis was conducted three times and yielded different LOD scores each time. Subsequent analyses were therefore run using 10,000 random samples.

4.5.5 Non-Parametric Linkage (NPL) Analysis using MERLIN

Multipoint NPL was performed using MERLIN, which uses the NPL_{pairs} and NPL_{all} statistics to score IBD allele sharing among pairs and groups of affected family members respectively (Whittemore & Halpern, 1994). To evaluate the evidence for linkage the (Kong & Cox, 1997) exponential model available in MERLIN was used. This model is offered as an alternative to the standard linear model, which is indicated for studies aiming to identify small increases in allele sharing among a large number of nuclear families. The exponential model is offered as a better option for studies aiming to identify large increases in allele sharing among a small number of families, and was therefore used in the BBF analyses. An allele sharing statistic (Delta) that measures IBD allele sharing between pairs and groups of affected family members in the NPL_{pairs} and NPL_{all} tests is reported in NPL tests.

In all of the MERLIN analyses non-informative family members were dropped from the analysis using the --trim option. This was done to avoid 'dragging down' the linkage test statistic that occurs when relative pairs whose IBD allele sharing is uncertain or incomplete are kept in the analysis. IBD sharing scores are calculated under the null hypothesis of no linkage and inclusion of uninformative relative pairs that contribute to the test statistic biases the results by diluting the data set and therefore affecting the evidence of linkage it provides (Schork & Greenwood, 2004). Simulation studies have revealed that even a small percentage of uninformative pairs cause the NPL statistic in MERLIN (as well as Genehunter and Allegro) to be considerably reduced (Cordell, 2004).

NPL_{pairs} and NPL_{all} analyses were performed separately on 18 subfamilies from the BBF branches, and 12 subfamilies from Branch 1 using the narrow, broad,

super, and depression phenotype models. The per family option (using the –per Family flag) was used to get LOD scores for each subfamily as well as total LOD scores for all subfamilies.

4.56 X-Chromosome Analyses

MERLIN in X (MINX) was used to test for linkage in the X-chromosome. There are no manuscripts or published works describing MINX performance and algorithms, so little information is provided about it here. X-linked traits are often recessive with full penetrance in hemizygous males. In recessive X-linked diseases, the only affected people are females who are homozygous for the disease allele and males who are hemizygous with the disease allele. The only way a female can be homozygous, however, is if she receives a disease alleles from her affected father. In MINX males are scored as homozygous in the pedigree input file when perfuming linkage testing on the X-chromosome. NPL MINX analyses were conducted on Branch 1 and the BBF using 230 SNPs available on the X-chromosome after data cleaning under the narrow, broad, super, and depression phenotype models.

4.5.7 Multiple Testing

In linkage studies multiple tests are usually performed in an attempt to find linkage between a marker(s) and a disease susceptibility locus. In the BBF four different phenotype models defining who is affected and who is unaffected were used and several different analysis methods and inheritance models were applied to the data. With the large number of tests performed the probability of false positives, i.e. significant results without there being linkage occurring increases. Therefore, to correct for multiple testing the formula below for a corrected LOD score *(Z_C)* was used

$$Z_c = Z + Log_{10}(m)$$

Where *Z* is the LOD score threshold if a single tests were used and *m* is the total number of independent tests performed (Ott, 1999). To apply this formula for multiple testing corrections, the number of independent tests needed to be determined given that the statistical tests and phenotype models used in the linkage analyses were nested. The Matrix Spectral Decomposition method implemented in the matSpD software (Li & Ji, 2005) was used to estimate the equivalent number of independent tests performed under the parametric and non-parametric linkage analyses separately. This was done with the view that the two tests represent alternative hypotheses as to the properties of the underlying genetic mutations in the family.

Testing for linkage between disease and multiple markers also represents a case of multiple testing. However, correction for the number of SNPs tested was not performed because the false positive rate due to the large number of SNPs is counterbalanced by the increase in the prior probability of linkage, as more and more

of the genome is excluded with the testing of multiple markers (Terwilliger & Ott, 1994).

4.5.7.1 Data Simulations

Data simulation is an alternative approach used to estimate empirical genome wide suggestive and significant levels. An average of 500 to 1000 simulations is usually generated under the assumption of no linkage. Random datasets that look like the original data in terms of marker informativeness, spacing and missing data patterns are generated while the pedigree structure and phenotypic information is preserved. The pedigree files generated are then analysed using the same linkage statistics, phenotype and disease inheritance models utilised in the original data. LOD scores exceeding a pre-set threshold for suggestive and significant linkage are recorded for each simulated dataset and counted. These LOD scores represent the number of false positives. A suggestive threshold reflecting LOD scores occurring by chance once in a genome scan and a significant threshold reflecting LOD sores occurring once in 20 genome scans is set. Using this method to estimate significance levels was the first choice of this study. However, due to delays in the computing cluster and the long time the simulations were taking to finish, this method was abandoned.

Chapter 5 Whole Genome Linkage Results

After data cleaning *(chapter 4 section 4.4)* whole genome linkage analyses were separately conducted on two sets of data: Branch 1 of the Brazilian Bipolar Family (Branch 1) and Branches 1, 2, and 3 of the Brazilian Bipolar Family (BBF), using 5315 autosomal SNPs genotyped using the Affymetrix 10K array and four phenotype models; narrow, broad, super, and depression. Multipoint parametric and non-parametric linkage (NPL) analyses were performed using the linkage software MERLIN (Abecasis, Cherny, Cookson, & Cardon, 2002) on 12 subfamilies from Branch 1 (n=269) and 18 subfamilies from the BBF (n=400). In addition, multipoint parametric linkage analyses were performed using the linkage software McLinkage Thomas, Camp, Farnham, Allen‐Brady, & Cannon‐Albright, 2008) where Branch 1 (n=242) and the BBF (n=361) were separately analysed with their structures intact. Parametric Linkage analyses were conducted under dominant and recessive modes of disease transmission and NPL analyses were performed using NPL_{all} and NPL_{pairs} (Figure 5.1).

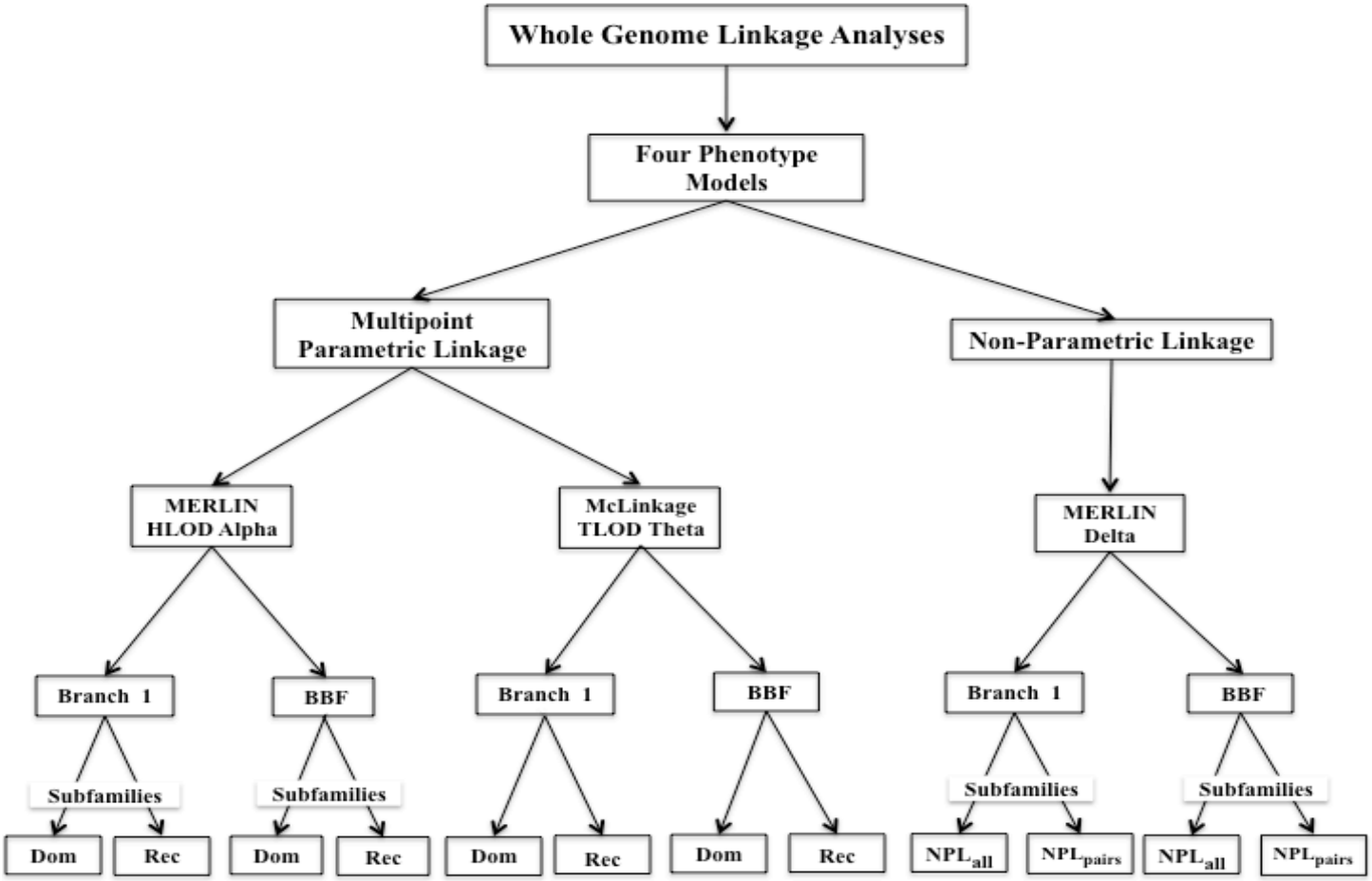


Figure 5.1. Flow chart detailing the whole genome parametric and non-parametric linkage analyses performed on Branch 1 and the BBF using four phenotype models; narrow including bipolar I, bipolar II, schizoaffective disorder/bipolar type; broad including all narrow phenotypes plus bipolar not otherwise specified and cyclothymia; super including all broad phenotypes plus one episode of major depressive disorder, recurrent episodes of major depressive disorder, and dysthymia and; depression including one episode of major depressive disorder, recurrent episodes of major depressive disorder, and dysthymia. The parametric linkage analyses were conducted under dominant (Dom) and recessive (Rec) modes of disease transmission. Non-parametric linkage was conducted using NPL_{all} and NPL_{pairs}.

5.1 Defining the Linkage Loci

5.1.1 Whole genome Linkage Thresholds

Following the recommendation of Morton (1955), a LOD score greater or equal to 3.0 and a LOD score greater or equal to 2.0 were used in the analyses to indicate genome wide significant and suggestive evidence for linkage respectively. Correction for multiple testing was applied using the formula by (OTT, 1999)

$$Z_c = Z + Log_{10}(m)$$

Where *Z* is the LOD score threshold if a single test was performed and *m* is the total number of independent tests performed. As the statistical tests and phenotype models used in the linkage analyses were nested, the matrix spectral decomposition method implemented in the matSpD software (Li & Ji, 2005) was used to estimate the equivalent number of independent tests performed under the parametric and NPL analyses separately. For the parametric linkage analyses, a spearman correlation matrix of sixteen variables including LOD scores achieved from the analyses performed on Branch 1 and the BBF using the four phenotype models under both the recessive and dominant modes of disease transmission (n=85040) was created in R, the statistical computing software (Team, 2010). The correlation matrix of the sixteen variables was entered into the matSpD program, which estimated the number of independent tests performed to be equivalent to eight *(Appendix I)*. Likewise, for the NPL analyses, a spearman correlation matrix of sixteen variables including LOD scores achieved from the NPL_{all} and NPL_{pairs} analyses performed on Branch 1 and the BBF under the four phenotype models (n=85040) was created in R. The correlation matrix of the sixteen variables was entered into the matSpD program, which estimated the number of independent tests performed to be equivalent to five *(Appendix II)*.

According to the number of independent tests estimated for the parametric and NPL analyses and using the formula by Ott (1999) to correct for multiple testing, a LOD score greater or equal to 3.9 and a LOD score greater or equal to 2.9 were used in the parametric linkage analyses to indicate whole genome significant and suggestive linkage respectively. A LOD score greater or equal to 3.7 and a LOD score greater or equal to 2.7 were used in the NPL analyses to indicate whole genome significant and suggestive linkage respectively.

5.1.2 Core and Maximum Linkage Regions

Linkage regions may be difficult to define, particularly in complex disease where broad flat peaks, or multiple peaks without clearly defined apexes are common. As such, chromosomal regions that achieved whole genome significant or suggestive linkage were defined in terms of a core and a maximum region. The core region was defined by markers with significant LOD scores (i.e. greater or equal to 3.9 or 3.7 for significance in parametric and NPL results respectively), or suggestive LOD scores (i.e. greater or equal to 2.9 and 2.7 for suggestive parametric and NPL results respectively), whereas the maximum region was defined by markers with a 1-LOD drop in threshold from the maximum LOD score achieved in the core region (Barnes, 2003).

5.2 Information Content

Information content mapping measures the total inheritance information extracted by the map of markers used in a study. Based on the marker density of the SNP set used in the analyses, Branch 1 subfamilies were estimated by MERLIN to have on average inheritance information of 81% with chromosome 22 having the least amount of information (76.7%). The information content across the BBF subfamilies was significantly lower than Branch 1 with an average of 65% across all chromosomes, with the lowest being chromosome 22, which captured the least amount of inheritance information (58.9%) (Figure 5.2). The reported sub-optimal coverage of the genome in Branch 1 and the BBF could result in false position linkage findings, significant linkage being missed, or inaccurately positioned on the genetic map. The information content of Branch 1 and the BBF with their structures in tact was not measured due to software limitations. However, one could assume that the information content of a connected pedigree is greater than that of a pedigree broken into subfamilies.

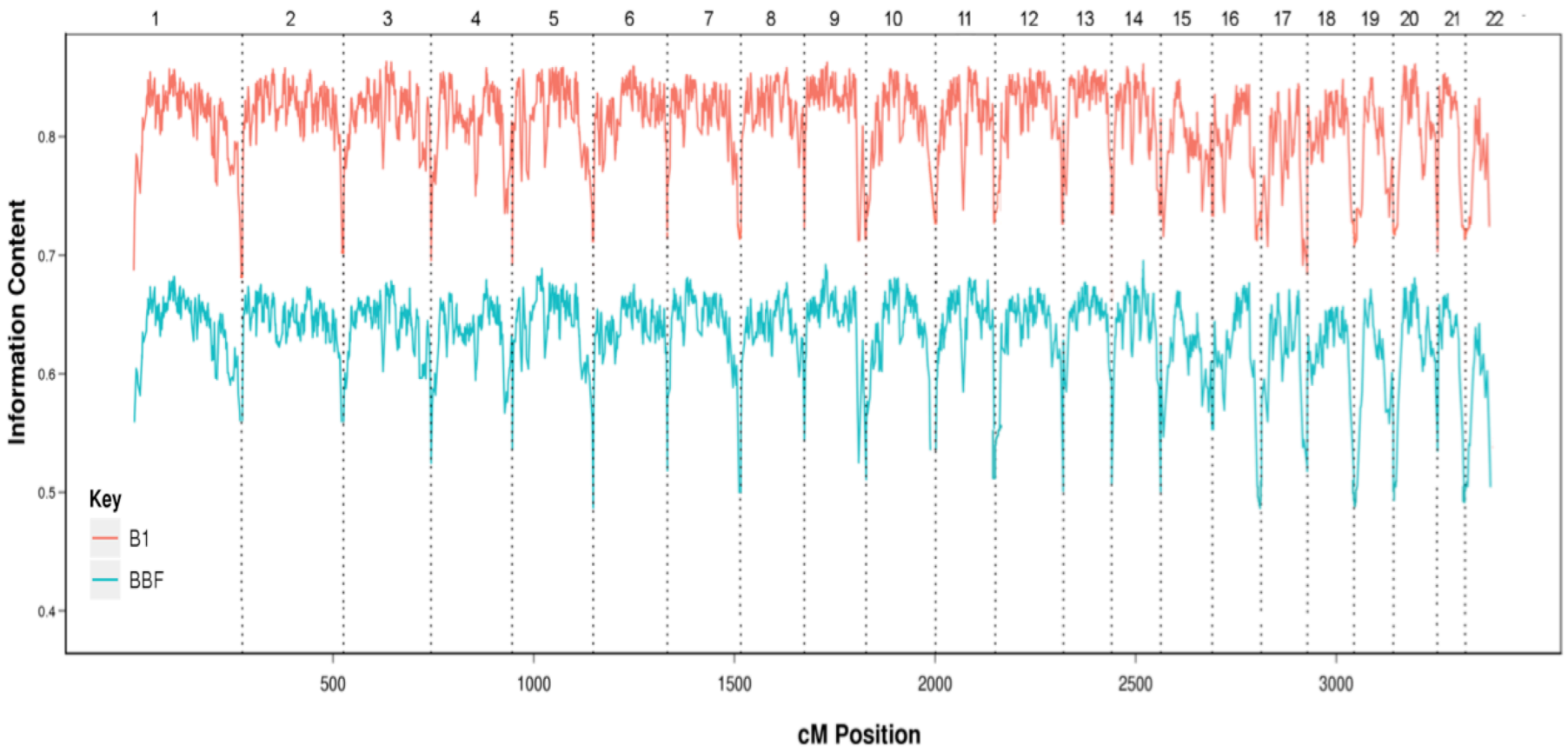


Figure 5.2. Information content of Branch 1 subfamilies (B1) and BBF subfamilies per chromosome as estimated by MERLIN entropy measure. The information content of Branch 1 subfamilies is significantly higher than that of the BBF subfamilies.

5.3 Summary of Whole-Genome Linkage Results

Using the corrected whole genome significant and suggestive thresholds, four chromosomal regions resulted in whole genome significant LOD scores: 2p23.1-p22.3, 3p24.3-p24.1, 11p15.4, and 12q24.22-q24.32 and four chromosomal regions resulted in whole genome suggestive LOD scores: 1p22.2-p21.3, 1q21.1-q21.3, 12p13.32-p13.31, and 22q11.21-q12.1 (Table 5.1). Graphs illustrating multipoint parametric and NPL peaks across the genome under the four phenotype models for Branch 1 and BBF subfamilies are presented in Figures 5.3 through 5.10. These graphs correspond to the MERLIN results only[1]. As seen in these graphs, MERLIN multipoint parametric and NPL analyses generally did not yield similar results, although some support across linkage methodologies was observed for the regions on chromosome 11p15.4 and 22q11.21-q12.1 with LOD scores greater or equal to 1.0. In addition, most of the linkage results were derived from Branch 1 subfamilies as linkage peaks either remained the same or decreased in size with the inclusion of subfamilies from Branches 2 or 3 in the BBF analyses. The exception is chromosome 1q21.1-q21.3, where a suggestive linkage peak (maximum LOD=2.83) was only achieved in analysis of the BBF subfamilies.

[1] Graphing the linkage results from McLinkage across the genome was unfeasible given that TLOD scores vary according to the recombination fraction theta.

Phenotype Model	Number Affected	Chromosomal Region	Software	Test	Mode Transmission	Maximum LOD
Narrow	40	1p22.2-p21.2	MERLIN	NPL_{all}	N/A	2.96*
		3p24.3-p24.1	McLinkage	TLOD	Dominant	4.18**
Broad	52	1p22.2-p21.3	MERLIN	NPL_{all}	N/A	2.93*
		22q11.21-q12.1	MERLIN	HLOD	Recessive	3.76*
Super	111	1q21.1-q21.3	MERLIN	NPL_{pairs}	N/A	2.83*
		2p23.1-p22.3	MERLIN	NPL_{pairs}	N/A	3.83**
Depression	59	11p15.4	MERLIN	NPL_{all}	N/A	4.49**
		12p13.32-p13.31	McLinkage	TLOD	Dominant	3.09*
		12q24.22-q24.32	McLinkage	TLOD	Dominant	4.74**

Table 5.1 Regions included in this table are those with significant or suggestive genome wide linkage in any analytic configuration. Two asterisks denote SNPs with whole-genome significance and one asterisk denotes SNPs with whole-genome suggestive evidence for linkage. All of the maximum LOD scores presented in the table are from analyses conducted on Branch 1, except the maximum LOD score for chromosomal region 1q21.1-q21.3 is from analysis conducted on the BBF.

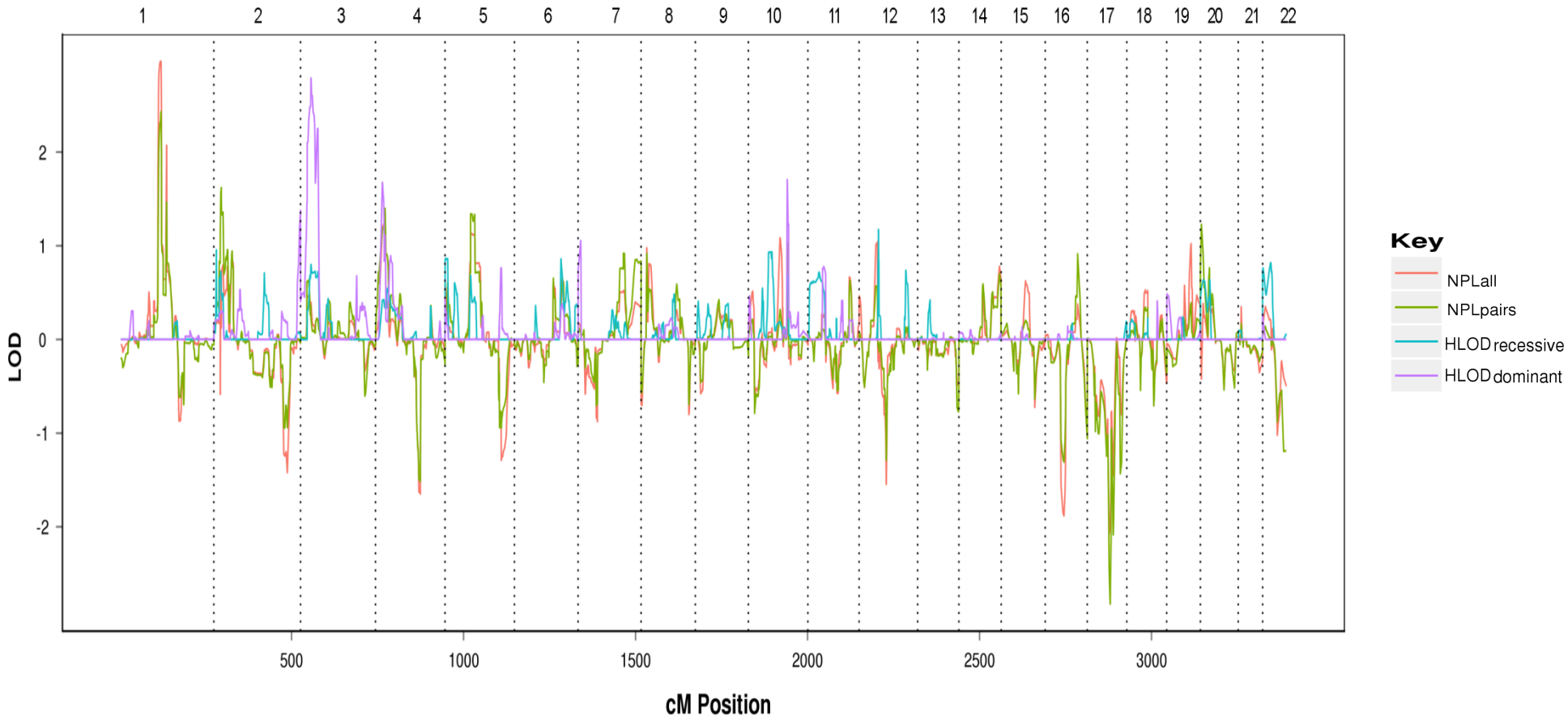


Figure 5.3. Parametric HLOD and non-parametric LOD scores for Branch 1 subfamilies under the narrow phenotype model. Chromosomes are separated by dotted lines. Of note the long regions of negative NPL LOD scores observed on chromosome 17 are due to a reason unknown to us and persist despite stringent quality control (Usually long stretches of negative NPL LOD scores calculated in MERLIN reflect pedigree or genotyping errors).

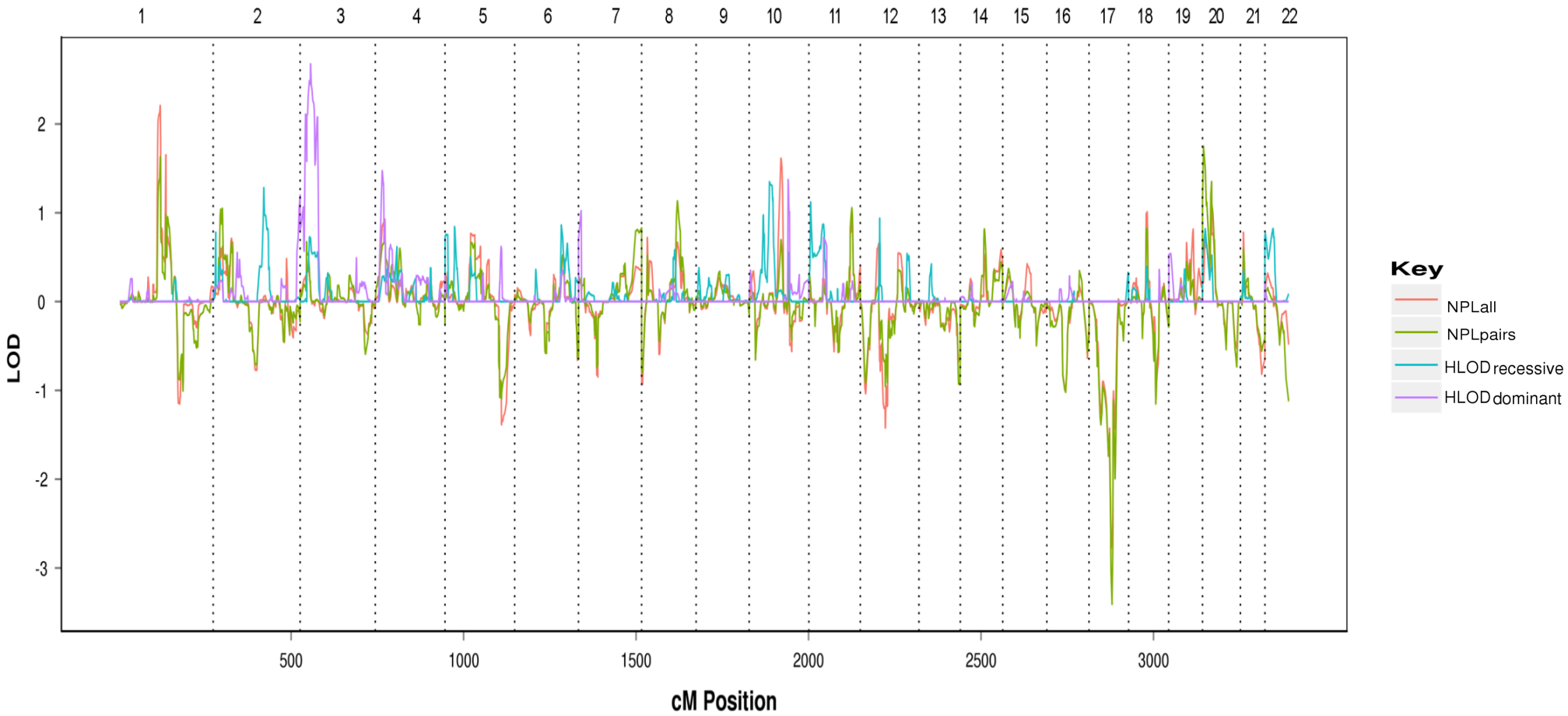


Figure 5.4. Parametric HLOD and non-parametric LOD scores for the BBF subfamilies under the narrow phenotype model. Chromosomes are separated by dotted lines. Of note the long regions of negative NPL LOD scores observed on chromosome 17 are due to an unknown reason to us and persist despite stringent quality control. (Usually long stretches of negative NPL LOD scores calculated in MERLIN reflect pedigree or genotyping errors)

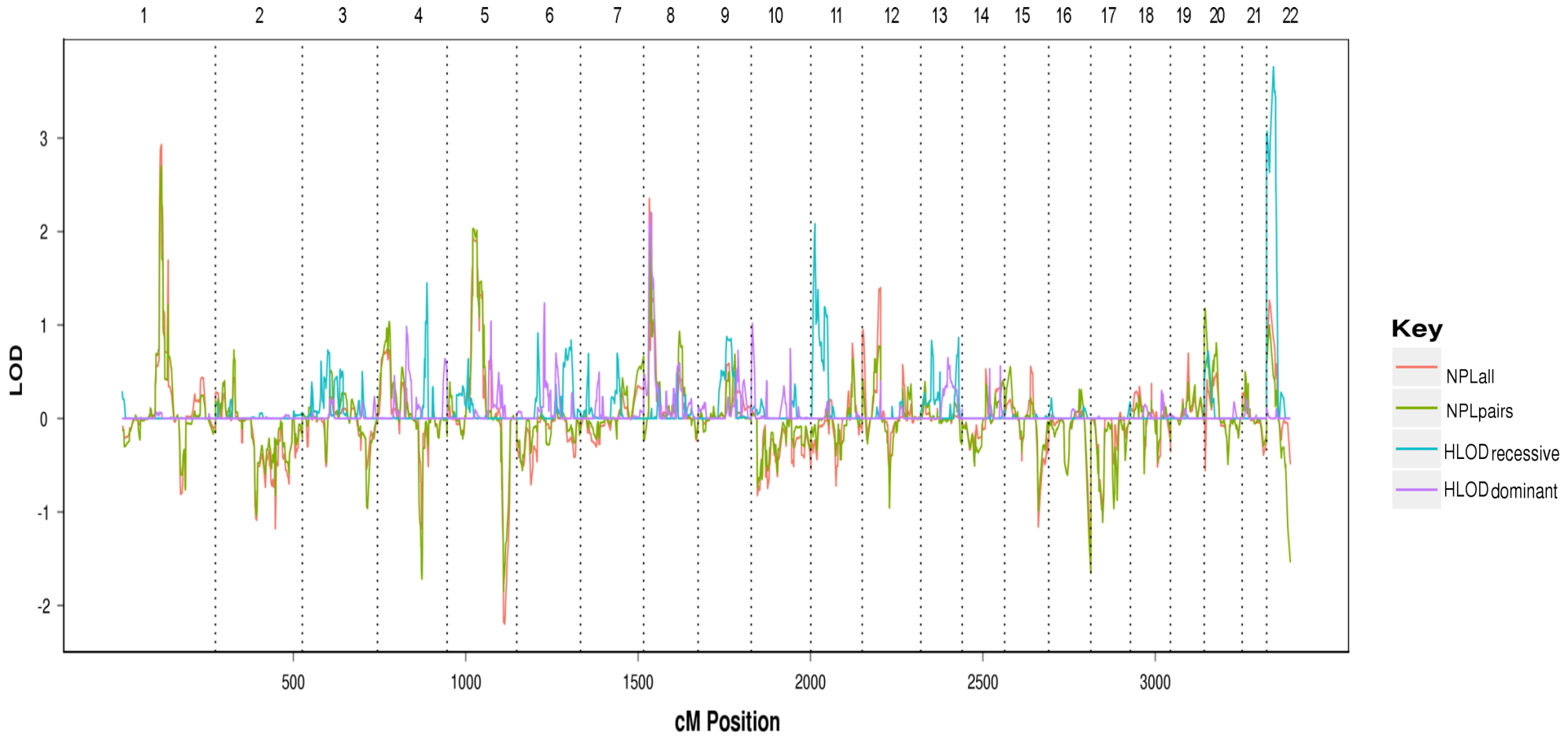


Figure 5.5. Parametric HLOD and non-parametric LOD scores for Branch 1 subfamilies under the broad phenotype model. Chromosomes are separated by dotted lines.

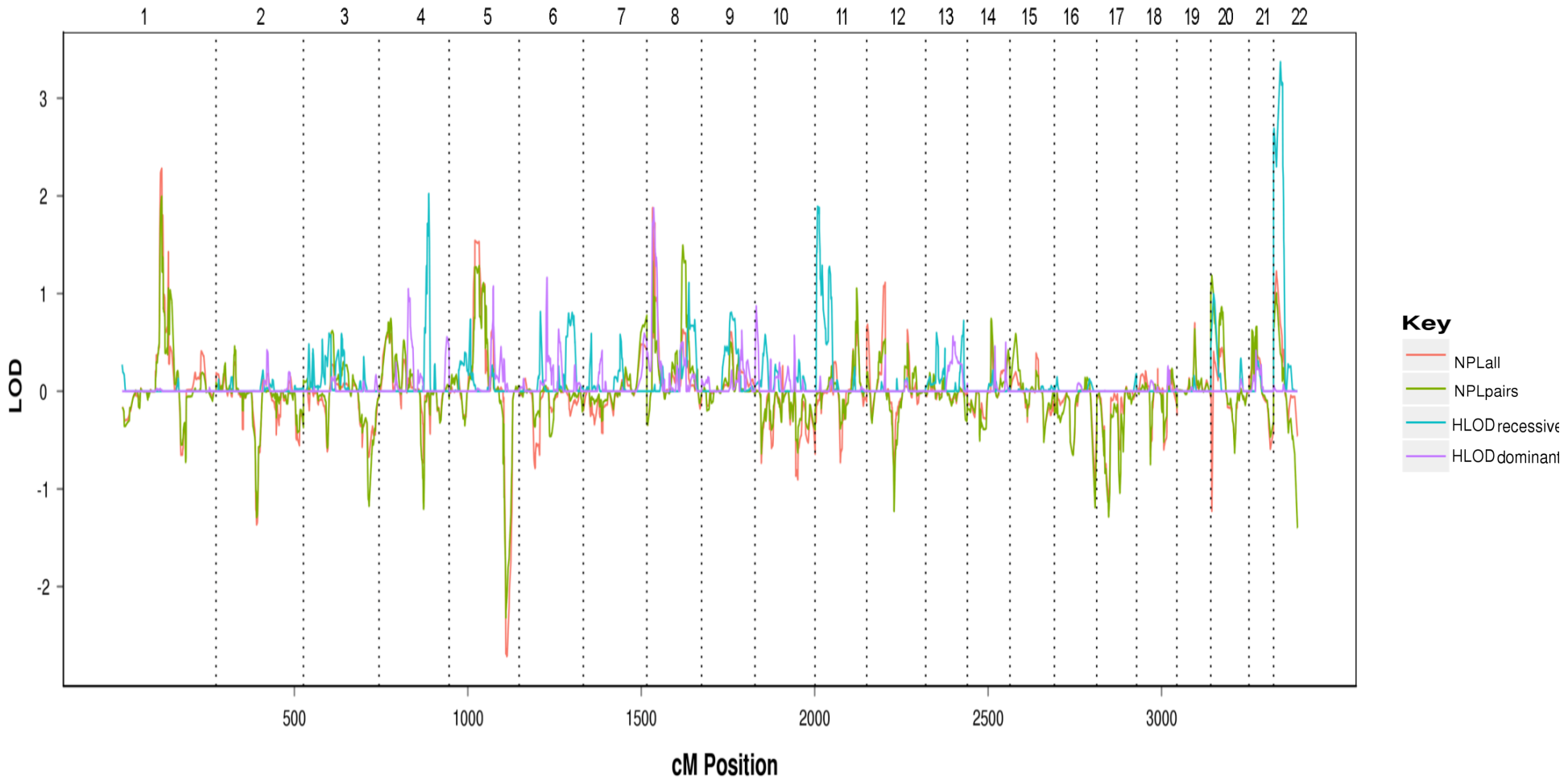


Figure 5.6. Parametric HLOD and non-parametric LOD scores for the BBF subfamilies under the broad phenotype model. Chromosomes are separated by dotted lines.

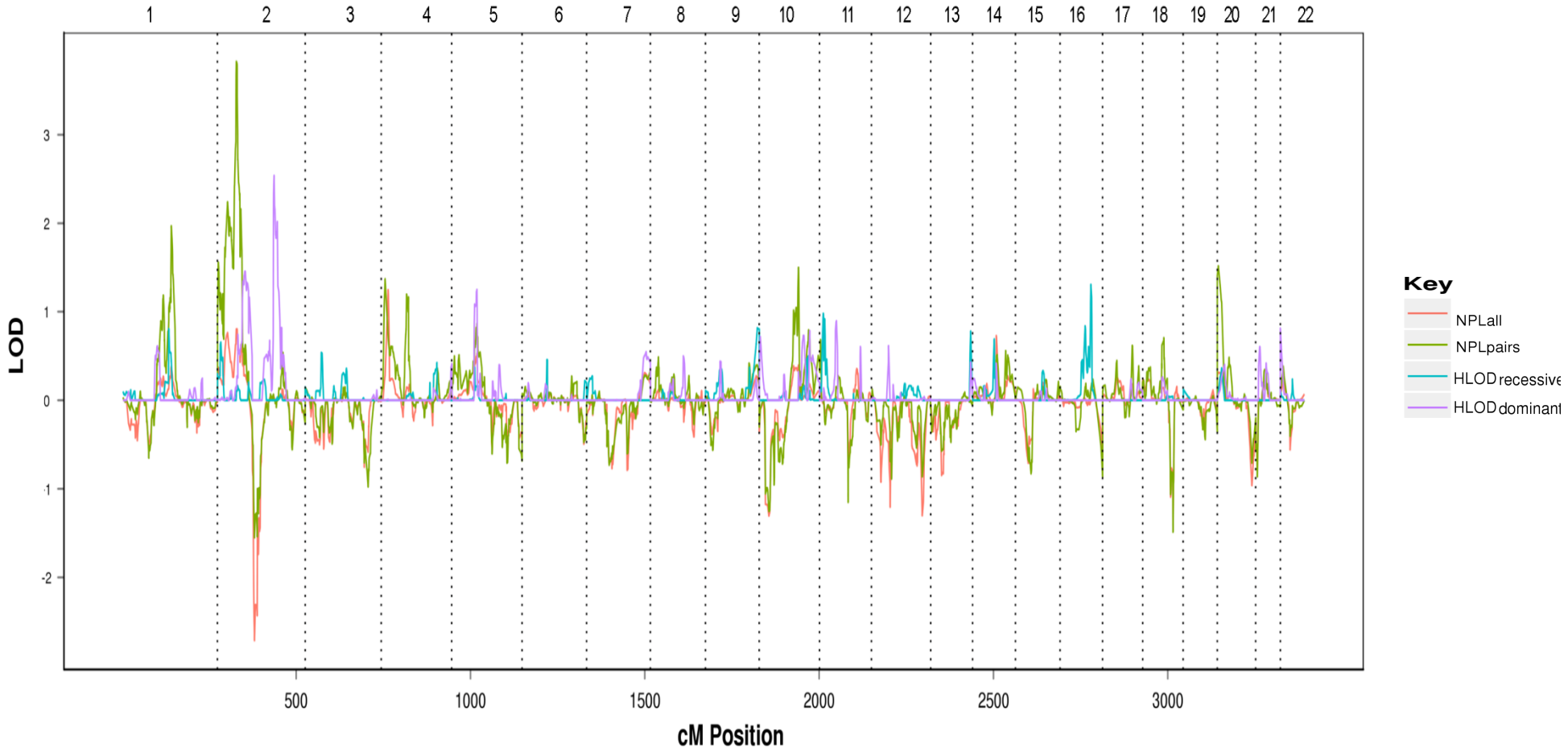


Figure 5.7. Parametric HLOD and non-parametric LOD scores for Branch 1 subfamilies under the super phenotype model. Chromosomes are separated by dotted lines.

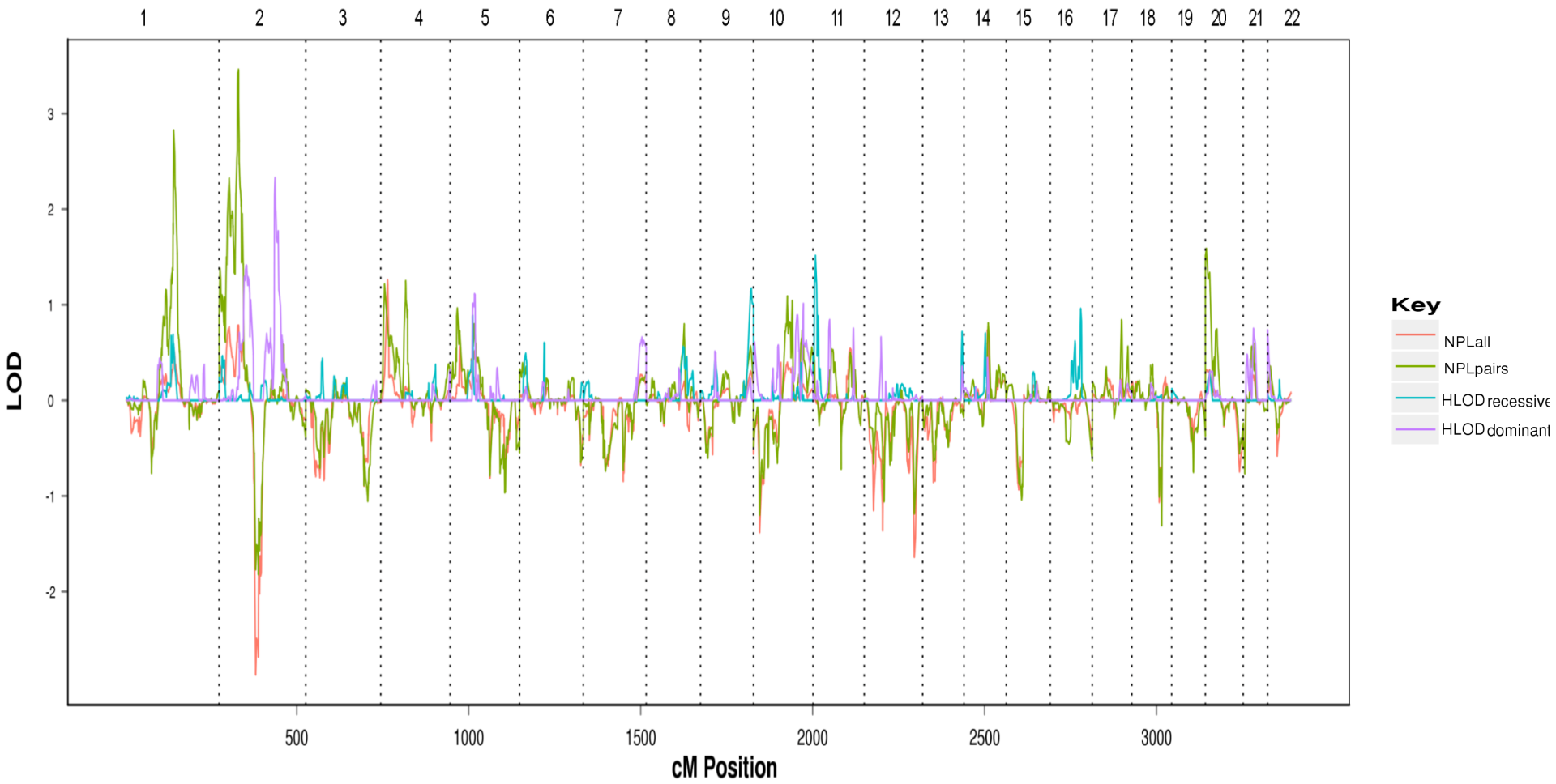


Figure 5.8. Parametric HLOD and non-parametric LOD scores for the BBF subfamilies under the super phenotype model. Chromosomes are separated by dotted lines.

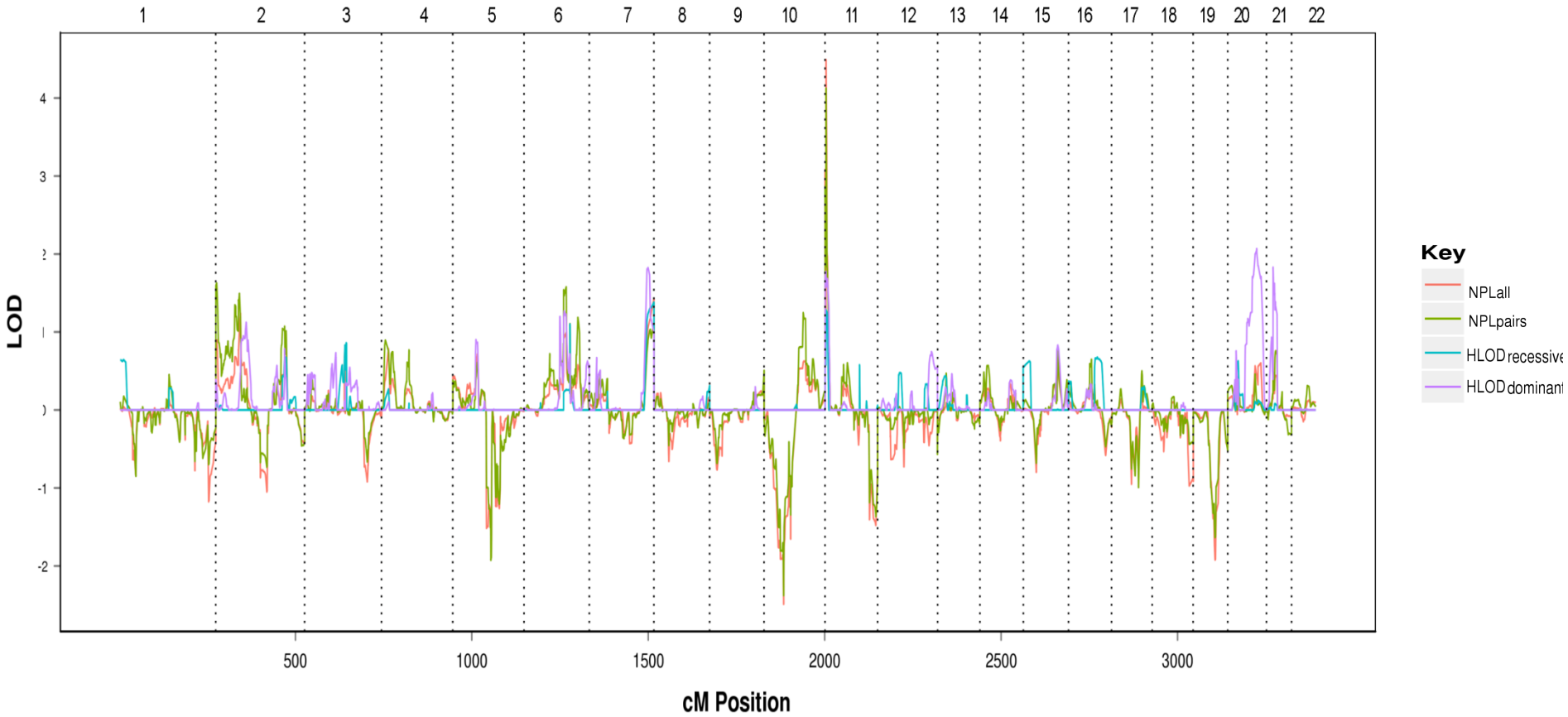


Figure 5.9. Parametric HLOD and non-parametric LOD scores for Branch 1 subfamilies under the depression phenotype model. Chromosomes are separated by dotted lines.

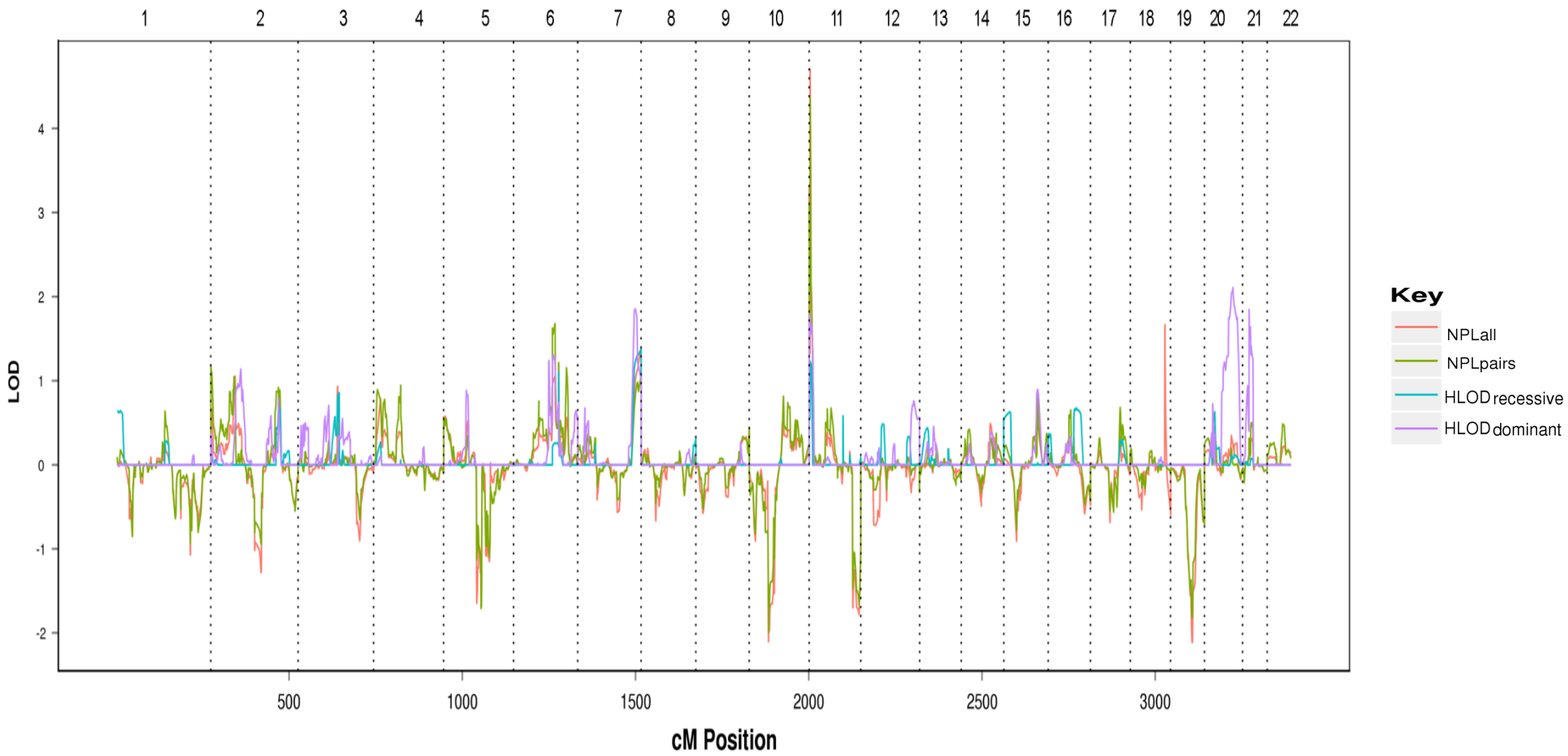


Figure 5.10. Parametric HLOD and non-parametric LOD scores for the BBF subfamilies under the depression phenotype model. Chromosomes are separated by dotted lines.

5.4 Parametric Linkage Results

This section will commence with a brief summary of the statistics employed by MERLIN and McLinkage in the analyses with the aim of clarifying their use in the study. Chromosomal regions achieving the corrected LOD scores for significant (LOD scores greater or equal to 3.9) and suggestive (LOD scores greater or equal to 2.9) linkage will then be presented per chromosome.

5.4.1 Parametric Linkage Statistics

Parametric linkage using MERLIN performs a LOD score test as well as heterogeneity LOD (HLOD) score test that accounts for the possibility of a mixture of linked and unlinked families to the marker(s) under investigation. An admixture parameter alpha (α) is incorporated into the HLOD score test, and represents the proportion of families with evidence for linkage. Of note is that α does not translate into the fraction of families with positive and negative LOD scores (Abecasis 2011, personal communication, 2011, *see Appendix III*). I interpreted this to be because the exclusion of linkage only occurs at a LOD score smaller or equal to -2, so families with LOD scores ranging from zero to -1.99 are considered not to have significant evidence against linkage to the marker(s) under investigation.

In the presence of locus heterogeneity, the reported HLOD scores are usually higher than the reported LOD scores. In our analyses, the HLOD scores were in all cases higher than the LOD scores in significant or suggestive regions indicating locus heterogeneity (Badner, Gershon, & Goldin, 1998; Morissette et al., 1999). In fact none of the LOD scores reported from the MERLIN analyses were greater than 1.0 (A

LOD score of 1 is arbitrary and corresponds to a p-value of 0.02[2], however, it is often reported in the literature to signify regions that warrant further investigation). As such, only the HLOD scores and associated α parameters are reported in this section.

Parametric linkage analysis using McLinkage performs a TLOD score test, which uses a multipoint approach to evaluate evidence for linkage over multiple markers and then uses a two point approach to calculate the LOD score at each marker by allowing values of the recombination fraction theta (θ) to vary (Abkevich, et al., 2001) *(see Chapter 4 section 4.5.4.6)*. Values of theta vary from zero, which indicates tight linkage, to 0.5, which is equivalent to independent assortment, or the null hypothesis of no linkage. In terms of distance between the marker and the putative disease locus, a theta of zero estimates that no genetic distance separates the two loci, a theta of 0.01 estimates the distance between the disease and the marker locus to be approximately 1 cM, and a theta of 0.02 estimates a distance of 2 cM. When theta is small, the genetic distance and theta are approximately equal. In practice geneticists treat them as equal for theta values of 0.1 or less. To achieve accurate genetic distances for theta values greater than 0.1 the Kosambi function *(see Chapter 4section 4.5.4.4)* was used. We will now explore significant and suggestive parametric linkage results per chromosome.

[2] The LOD score was converted to a p-value by multiplying by 4.6 for a chi-square and then taking the value at 1 degree of freedom and dividing it by 2 (OTT, 1999)

5.4.2 Chromosome 3p

5.4.2.1 The Narrow Phenotype Model: McLinkage Results

Branch 1 yielded a maximum TLOD score of 4.18 supportive of whole genome significant linkage on chromosome 3p24.3 in a core region of 518 kb using the narrow phenotype model under a dominant mode of disease transmission. This occurred at a recombination fraction of 0.2, which is indicative of "loose" linkage, as it estimates the disease and marker locus to be approximately 19 cM[*] away (Figure 5.11). A maximum region defined by a 1-TLOD drop in threshold from 4.18 identified an 8.4 Mb susceptibility region on chromosome 3p24.3-p24.1 (Table 5.2). Using the same phenotype model and mode of disease transmission, analysis of the BBF yielded a maximum TLOD score of 3.46 supportive of whole genome suggestive linkage in a core region spanning 3.3 Mb on chromosome 3p24.3-p24.2 at a recombination fraction of 0.2, indicating the disease and marker may be approximately 19 cM apart (Figure 5.11). Including markers with a 1-TLOD drop in threshold from 3.46 expanded the region to a maximum of 4.5 Mb. This region is 3.9 Mb smaller than the maximum region identified in Branch 1. A shift in linkage peaks is observed between the findings from Branch 1 and the BBF as the marker achieving the maximum TLOD score in the BBF (rs2045300) is approximately 3.2 Mb away from the closest marker (rs904827) achieving the maximum TLOD score in Branch 1. Further, it is questionable that markers that achieved TLOD scores greater than 3.0 in Branch 1 yielded negative TLOD scores in the BBF (e.g. rs2370990). Of note, however, is that the region identified in the BBF is only suggestive of linkage and therefore more prone to being a false positive.

[*] As theta is greater than 0.1 the Kosambi function M=1/2 tanh (2r) where M is the distance in Morgans and r is the recombination fraction or theta.

Using the narrow phenotype model, this region did not report any suggestive or significant TLOD scores under the recessive mode of disease transmission.

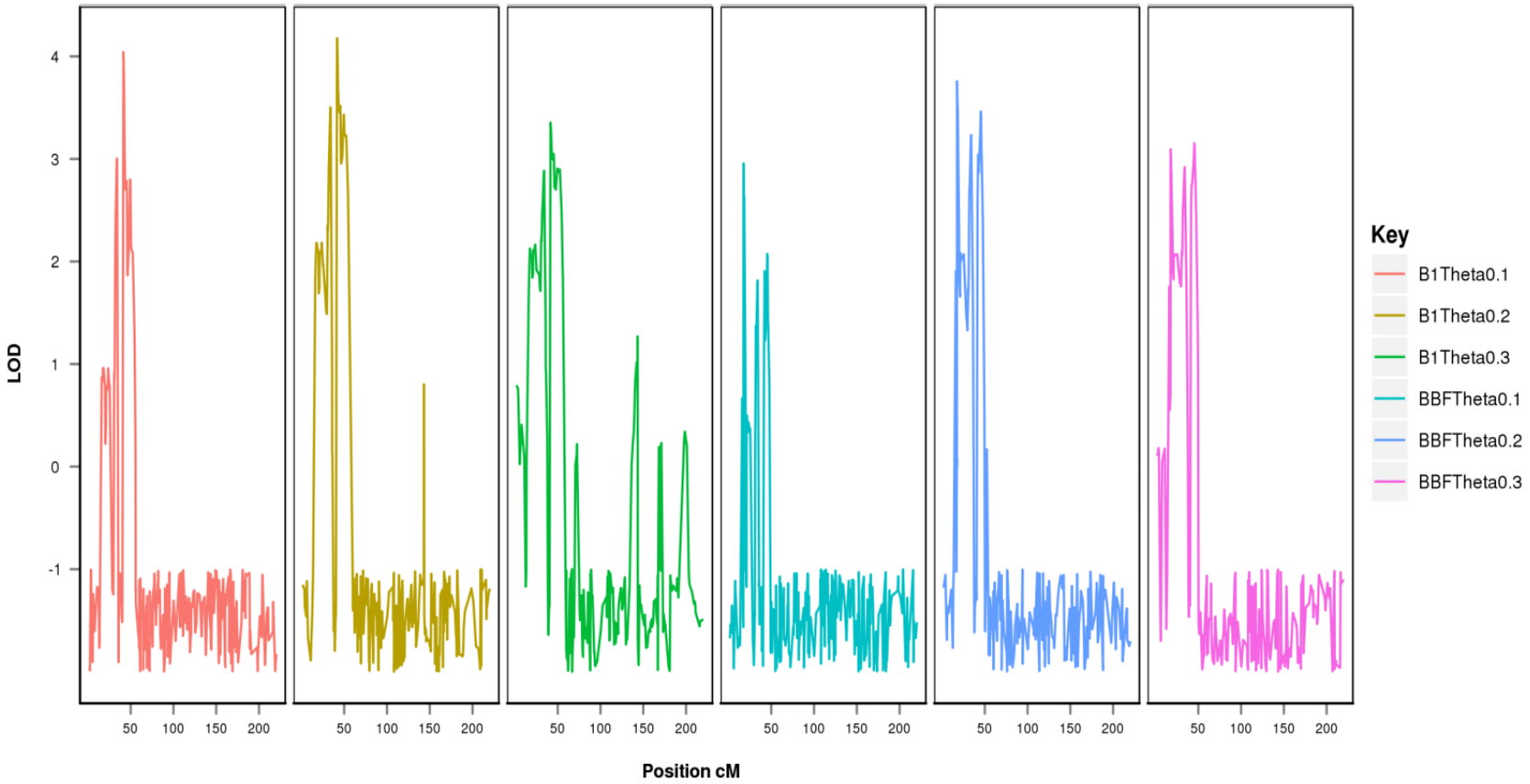


Figure 5.11. McLinkage TLOD scores for Branch 1 (B1) and the BBF at recombination fractions (Theta) ranging from 0.1 through to 0.3 on chromosome 3 versus centiMorgan position using the narrow phenotype model under the dominant mode of disease transmission. The TLOD scores in Branch 1 and the BBF were maximised at Theta=0.2. The maximum linkage peaks occur at different locations in Branch 1 and the BBF. TLOD scores smaller or equal to -2 were truncated in the graph for ease of presentation.

SNP	Physical Position	cM Position	Branch 1	BBF
rs1392573	20695652	41.56	4.18**	2.60
rs904827	20696379	41.56	4.18**	2.61
rs725542	21213394	42.22	4.02**	3.04*
rs721623	21650537	42.77	3.70*	2.90*
rs2358693	22574169	43.58	3.50*	2.99*
rs1074612	22579568	43.59	3.49*	3.00*
rs778480	23498492	44.75	3.45*	3.26*
rs2045300	23871402	45.34	3.52*	3.46*
rs951015	24053792	45.63	3.43*	3.41*
rs1112195	24085166	45.68	3.40*	3.42*
rs1394764	24405050	46.18	2.97*	3.05*
rs2196427	24533278	46.39	2.96*	3.03*
rs720822	25232155	47.77	3.02*	2.37*
rs1348979	27027497	49.44	3.43*	0.68
rs2037472	27070331	49.49	3.43*	0.64
rs724244	27242153	49.69	3.42*	0.22
rs2370990	27690531	50.22	3.25*	-1.32
rs2371121	27850725	50.40	3.23*	-1.61
rs1609729	28303111	51.13	3.22*	-0.15
rs1979113	29057186	52.44	3.23*	0.17

Table 5.2. McLinkage TLOD scores at a recombination fraction of 0.2 on chromosome 3p24.3-p24.1 under the narrow phenotype model and dominant mode of disease transmission. Two asterisks denote SNPs with whole genome significance defined as LOD scores greater or equal to 3.90 and one asterisk denotes SNPs with whole genome suggestive evidence for linkage defined as LOD scores greater or equal to 2.90.

5.4.2.2 The Narrow Phenotype Model: MERLIN Results

MERLIN analysis identified a maximum region that spans 18.3 Mb on chromosome 3p25.3-p24.1 with near suggestive evidence for linkage (maximum HLOD score of 2.79 in Branch 1 and 2.68 in the BBF) that encompasses the susceptibility locus identified by McLinkage under the same phenotype model and mode of disease transmission (Table 5.3). However, the linkage peaks identified by the two programs are approximately 7.4 Mb apart (Figure 5.12). This region is only presented here as a point of comparison between the two programs.

SNP	Physical Position	cM Position	Branch 1		BBF	
			HLOD	α	HLOD	α
rs59585	13501254	31.98	2.79	0.58	2.68	0.53
rs1368575	13831304	32.43	2.79	0.57	2.64	0.53
rs934448	13831352	32.44	2.79	0.57	2.64	0.50
rs1392573	20695652	41.56	2.38	0.47	2.23	0.42
rs904827	20696379	41.57	2.38	0.47	2.23	0.40
rs725542	21213394	42.22	2.33	0.45	2.18	0.40
rs721623	21650537	42.77	2.30	0.44	2.16	0.39
rs2358693	22574169	43.58	2.12	0.43	1.98	0.39
rs1074612	22579568	43.59	2.12	0.43	1.98	0.49
rs778480	23498492	44.75	1.67	0.41	1.54	0.37
rs2045300	23871402	45.34	1.70	0.43	1.56	0.38
rs951015	24053792	45.63	1.72	0.43	1.58	0.38
rs1112195	24085166	45.68	1.72	0.43	1.58	0.38
rs1394764	24405050	46.18	1.74	0.44	1.59	0.39
rs2196427	24533278	46.39	1.74	0.44	1.59	0.39
rs720822	25232155	47.77	1.94	0.47	1.77	0.42
rs1348979	27027497	49.44	2.12	0.49	1.94	0.44
rs2037472	27070331	49.49	2.12	0.49	1.95	0.44
rs724244	27242153	49.69	2.14	0.49	1.96	0.44
rs2370990	27690531	50.22	2.18	0.49	2.01	0.44
rs2371121	27850725	50.40	2.20	0.49	2.02	0.44
rs1609729	28303111	51.13	2.25	0.49	2.07	0.44
rs1979113	29057186	52.44	2.25	0.49	2.08	0.44

Table 5.3. MERLIN near suggestive HLOD scores on chromosome 3p25.3-p24.1 under the narrow phenotype model and dominant mode of disease transmission. Only SNPs with the maximum HLOD scores and those that overlap the McLinkage region are presented in this table. For the full table see *Appendix IV*.

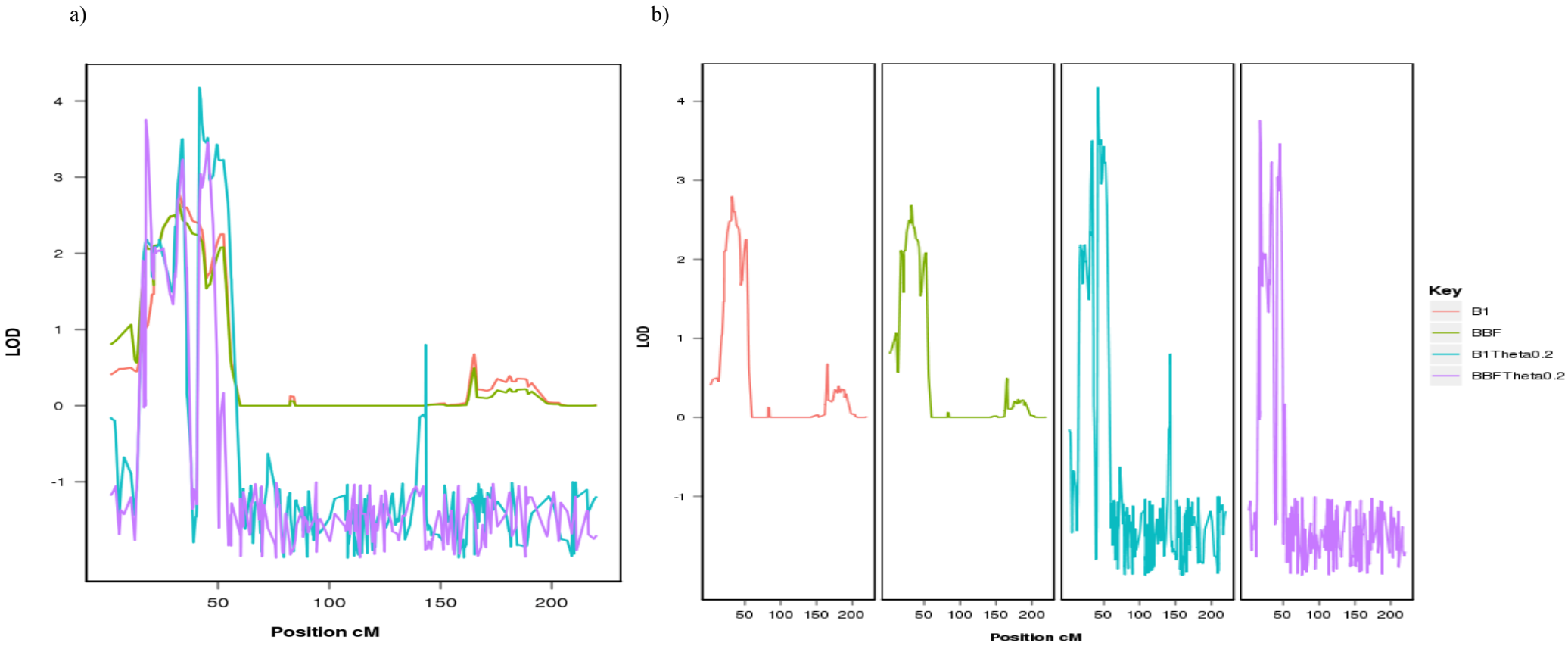


Figure 5.12. MERLIN HLOD and McLinkage TLOD scores (at recombination fraction of 0.2) for Branch 1 and the BBF on chromosome 3 versus centiMorgan position under the narrow phenotype model and dominant mode of disease transmission. McLinkage TLOD scores smaller or equal to -2 were truncated in the graph for ease of presentation. a) The MERLIN and McLinkage linkage regions overlap, however, the peaks occur 7.4 Mb apart. A clear shift in the linkage peaks defined by McLinkage in Branch 1 and the BBF is observed. b) Linkage graphs are presented separately for clearer viewing. The McLinkage peaks are not smooth and show great variation in TLOD scores within the linkage region.

5.4.2.2.1 Subfamily Contribution to the Heterogeneity LOD Score

The region on chromosome 3p24.3-p24.1 influences susceptibility to disorders defined under the narrow phenotype model (BPI, BPII, and SAD) in a proportion of subfamilies in Branch 1 and the BBF. Evaluation of subfamily contribution to the overall HLOD score in Branch 1 indicated that only six subfamilies contributed positively to the overall HLOD score. Subfamilies 2, 3, and 9 seemed to be driving the linkage signal in this region by contributing with nominally significant (p-values greater or equal to 0.05) HLOD scores of 1.48, 1.37, and 1.17[3]. Further, expanding the analysis to the BBF showed that in addition to the contributions of subfamilies from Branch 1 only subfamily 14 from Branch 2 contributed positively to the linkage signal in this region (Figure 5.13). These findings may be suggestive of heterogeneity at this locus.

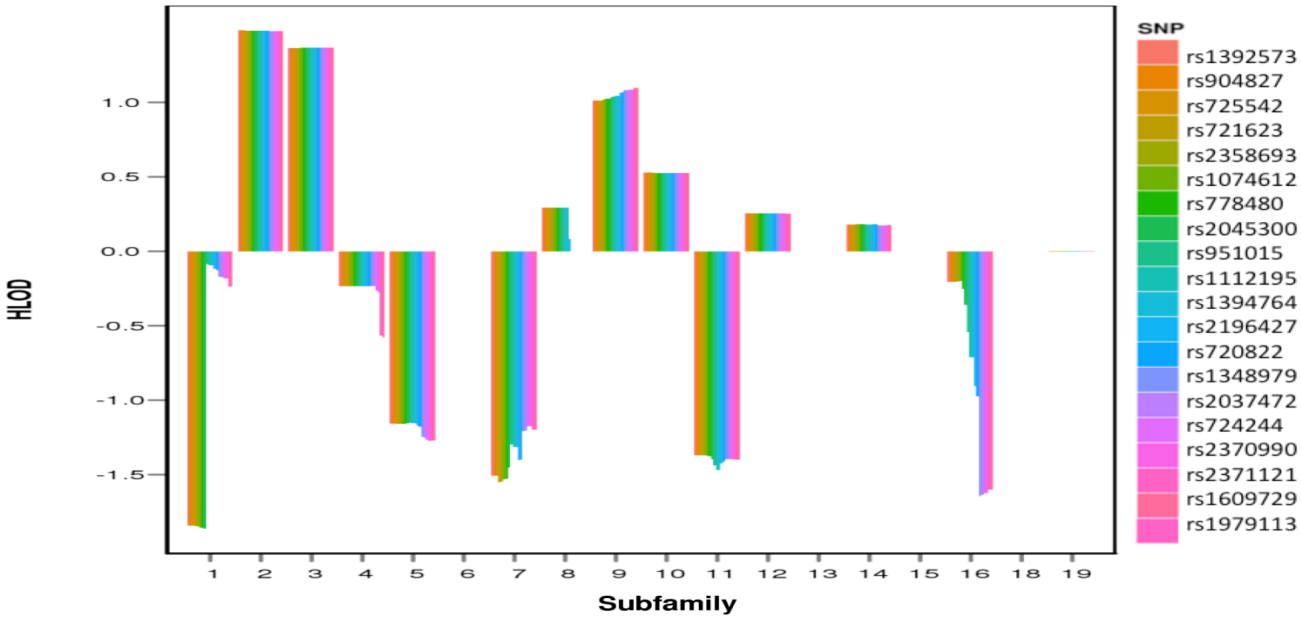


Figure 5.13. Subfamily contribution to the HLOD scores on chromosome 3p24.3-p24.1 using the narrow phenotype model under a dominant mode of disease transmission. SNPs in this region nearly achieved significance using MERLIN and were reported as significant using McLinkage (For family contribution to the HLOD score in the larger region, which showed a similar pattern, see *Appendix V*). Subfamilies 1 through to 12 constitute Branch 1, subfamilies 13 and 14 constitute Branch 2 and subfamilies 15 through to 19 constitute Branch 3. Subfamily 17 was dropped from the analyses.

[3] A LOD score of 0.59 is equivalent to a p-value of 0.05, which is nominally significant.

5.4.2.3 The Broad, Super and Depression Phenotype Models

The region identified on chromosome 3p24.3-p24.1 (and 3p25.3-p24.1) using the narrow phenotype model yielded no suggestive or significant evidence for linkage using the broad, super or depression phenotype models in MERLIN or McLinkage analyses under the dominant or recessive modes of disease transmission in Branch 1 or the BBF. These results implicate chromosome 3p24.3-p24.1, possibly 3p25.3-p24.1, as a susceptibility locus for phenotypes identified under the narrow model (BPI, BPII, and SAD) only.

5.4.3 Chromosome 12p

5.4.3.1 The Depression Phenotype Model: McLinkage Results

Assuming a dominant mode of disease transmission and using the depression phenotype model, Branch 1 yielded a maximum TLOD score of 3.09 supportive of whole genome suggestive linkage on chromosome 12p13.31 in a core region spanning 18.2 kb. This occurred at a recombination fraction of 0.1, which indicates that the disease and marker locus are approximately 10 cM apart. A maximum region defined by a 1-TLOD drop in threshold from 3.09 identified a 1.6 Mb disease susceptibility region on chromosome 12p13.32-p13.31 (Figure 5.14 in *section 5.4.4*). The TLOD score was reduced to a non-significant score of 2.02 occurring at a recombination fraction of 0.1 when the BBF was analysed (Table 5.4).

Using the depression phenotype model, this region did not report any suggestive or significant TLOD scores under the recessive mode of disease transmission

SNP	Physical Position	cM Position	Branch 1	BBF
Rs720334	4493794	13.22	2.31	0.95
Rs501004	5238215	14.79	2.35	1.08
Rs2041428	5375028	15.03	2.36	1.10
Rs1005221	5493117	15.14	2.36	1.10
Rs764220	5711247	15.63	2.33	1.11
Rs2192190	5741549	15.72	2.74	1.37
Rs723188	6063905	16.19	3.07*	2.02
Rs917858	6081938	16.207	3.09*	2.03
Rs917859	6082113	16.208	3.09*	2.03
Rs717180	9395807	24.008	-0.52	0.25
Rs252027	9412418	24.013	-0.52	0.25
Rs252028	9412883	24.014	-0.52	0.25

Table 5.4. McLinkage TLOD scores at a recombination fraction of 0.1 on chromosome 12p13.32-p13.31 under the depression phenotype model and a dominant mode of disease transmission. One asterisk denotes SNPs with whole genome suggestive evidence for linkage defined as LOD scores greater or equal to 2.90.

5.4.3.2 The Depression Phenotype Model: MERLIN Results

Using MERLIN, the region on chromosome 12p13.32-p13.31 reported no suggestive or significant linkage signals in either Branch 1 or the BBF using the depression phenotype model under a dominant or recessive mode of disease transmission.

5.4.3.3 The Narrow, Broad, and Super Phenotype Models

The linkage region identified on chromosome 12p13.32-p13.31 using the depression phenotype model yielded no suggestive or significant linkage signals using the narrow, broad or super phenotype models under the dominant or recessive modes of disease transmission in Branch 1 or the BBF. These results implicate chromosome 12p13.32-p13.31 as a susceptibility locus for phenotypes identified under the depression model (MDD and dysthymia) only.

5.4.4 Chromosome 12q

5.4.4.1 The Depression Phenotype Model: McLinkage Results

The highest linkage signal achieved in the whole genome linkage scan occurred on chromosome 12q24.31-q24.32 under a depression only phenotype model assuming a dominant mode of disease transmission. McLinkage identified a linkage peak in Branch 1 with a maximum TLOD score of 4.74 in a region spanning 120 kb at a recombination fraction of 0.01, which estimates the disease and marker locus are approximately 1 cM away. Surprisingly, the linkage peak is followed by a sharp drop in TLOD scores and a maximum region is therefore not defined. Analysis of the BBF reduced the linkage signal to a maximum TLOD score of 3.16 at recombination fraction 0.01, which only meets study criteria for whole genome suggestive linkage. A smoother decline in LOD scores is observed in the BBF analysis and a maximum region defined by a 1-TLOD drop in threshold from this maximum identified a 2.5 Mb susceptibility region on chromosome 12q24.31-q24.32. (Table 5.5, Figure 5.14).

Using the depression phenotype model, this region did not report any suggestive or significant TLOD scores under the recessive mode of disease transmission.

SNP	Physical Position	cM Position	Branch 1	BBF
rs325095	125885304	145.80	4.04**	3.06*
rs849346	126005019	146.256	4.74**	3.16*
rs861070	126005488	146.258	4.74**	3.16*
rs722704	126921171	149.21	2.28	3.51*
rs950591	127672428	151.60	0.92	2.71
rs3851661	127724123	151.77	0.13	2.67
rs721605	127980028	152.70	2.92*	2.69
rs1386848	128426363	154.36	2.44	2.19

Table 5.5. McLinkage TLOD scores at a recombination fraction of 0.01 on chromosome 12q24.22-q24.32 under the depression phenotype model and a dominant mode of disease transmission. Two asterisks denote SNPs with whole genome significance defined as LOD scores greater or equal to 3.90 and one asterisk denotes SNPs with whole-genome suggestive evidence for linkage defined as LOD scores greater or equal to 2.90.

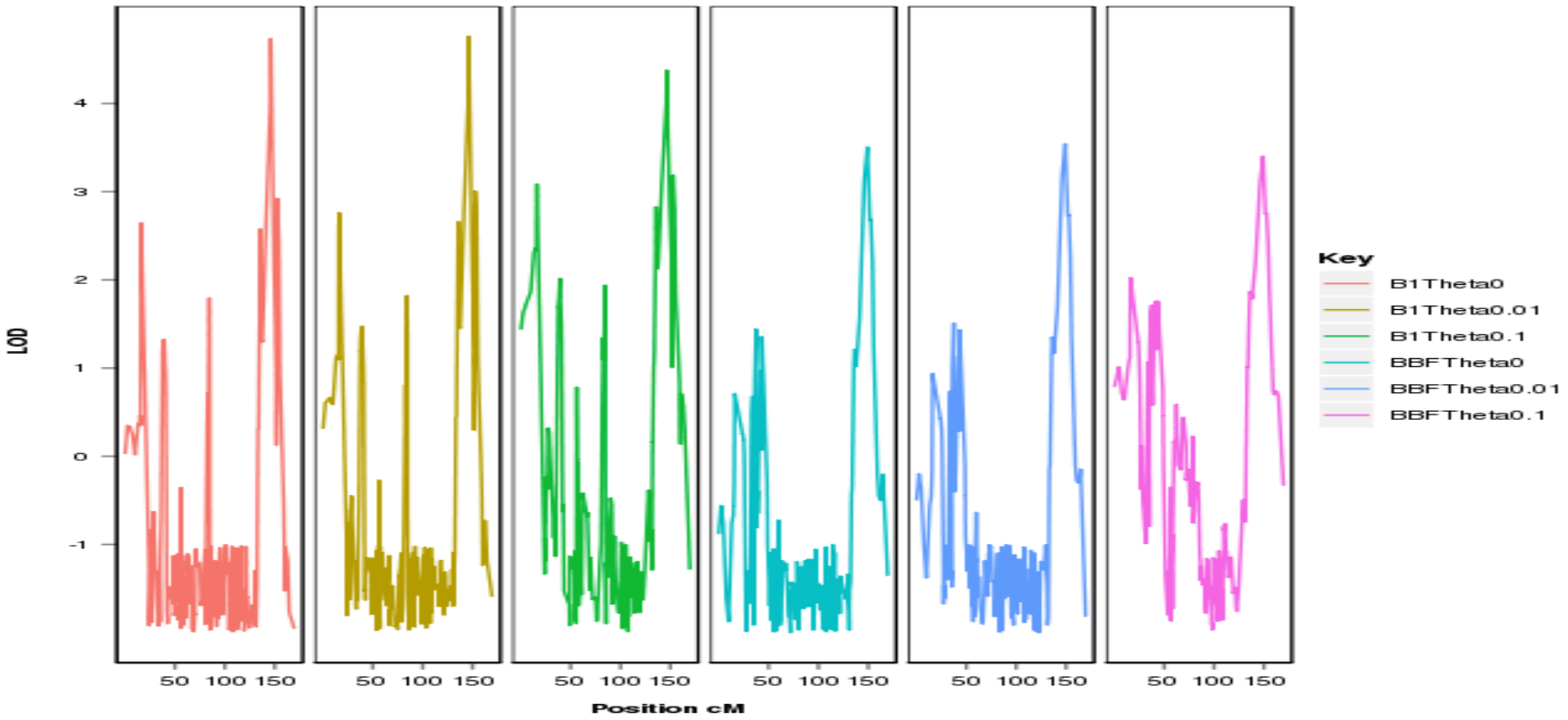


Figure 5.14. McLinkage TLOD scores for Branch 1 (B1) and the BBF at recombination fractions (Theta) ranging from 0.0 through to 0.1 on chromosome 12 versus centiMorgan position under the depression phenotype model. Two linkage peaks are observed on the two arms of chromosome 12, a suggestive linkage peak on chromosome 12p13.32-p13.31 maximising the TLOD scores at recombination fraction of 0.1, and a significant peak on chromosome 12q24.31-q24.32 maximising the TLOD scores at recombination fraction of 0.01. TLOD scores smaller or equal to -2 were truncated in the graph for ease of presentation.

5.4.4.2 The Depression Phenotype Model: MERLIN Results

MERLIN analyses using the depression phenotype model found no suggestive or significant linkage signals on chromosome 12q24.22-q24.32 under a dominant or recessive mode of disease transmission in either Branch 1 or the BBF, reporting LOD scores approaching zero.

5.4.4.3 The Narrow, Broad, and Super Phenotype Models

The linkage region identified on chromosome 12q24.22-q24.32 using the depression phenotype model revealed no suggestive or significant evidence for linkage using the narrow, broad, or super phenotype models in MERLIN or McLinkage analyses under the dominant or recessive modes of disease transmission in Branch 1 or the BBF. However, McLinkage TLOD scores greater than 1.0 were found at a recombination fraction of 0.2 under the super phenotype model (BPI, BPII, SAD, BPNOS, cyclothymia, MDD, and dysthymia) using the dominant mode of disease transmission in Branch 1 and the BBF. All of the other phenotype models had LOD scores of approximately zero. These results implicate chromosome 12q24.22-q24.32 as a susceptibility locus for phenotypes identified under the depression model (MDD and dysthymia) only.

5.4.5Chromosome 22q

5.4.5.1 The Broad Phenotype Model: McLinkage Results

McLinkage analysis of Branch 1 yielded a maximum TLOD score of 3.53 supportive of whole genome suggestive linkage on chromosome 22q11.21-q11.22 in a core region of 720 kb using the broad phenotype model under a recessive mode of disease transmission. This occurred at a recombination fraction of zero, which is indicative of tight linkage (no recombination). A maximum region defined by a 1-TLOD drop in threshold from 3.53 identified a 9.3 Mb susceptibility region on chromosome 22q11.21-q12.1. Using the same phenotype model and mode disease transmission, analysis of the BBF yielded a maximum TLOD score of 3.34 also supportive of whole genome suggestive linkage in a core region spanning 4.5 Mb on chromosome 22q11.21-q11.22. This evidence for linkage occurs at a recombination fraction of 0.1 estimating the disease and marker locus to be approximately 10 cM away, as opposed to the tight linkage identified in Branch 1 at a recombination fraction of zero. The maximum region defined in the BBF analysis by including markers with a 1-TLOD drop in threshold from 3.34 expanded this region to a maximum region of 9.7 Mb that encompasses the region identified in Branch 1 (Figure 5.15, Table 5.6).

Using the broad phenotype model, this region did not report any suggestive or significant TLOD scores under the dominant mode of disease transmission.

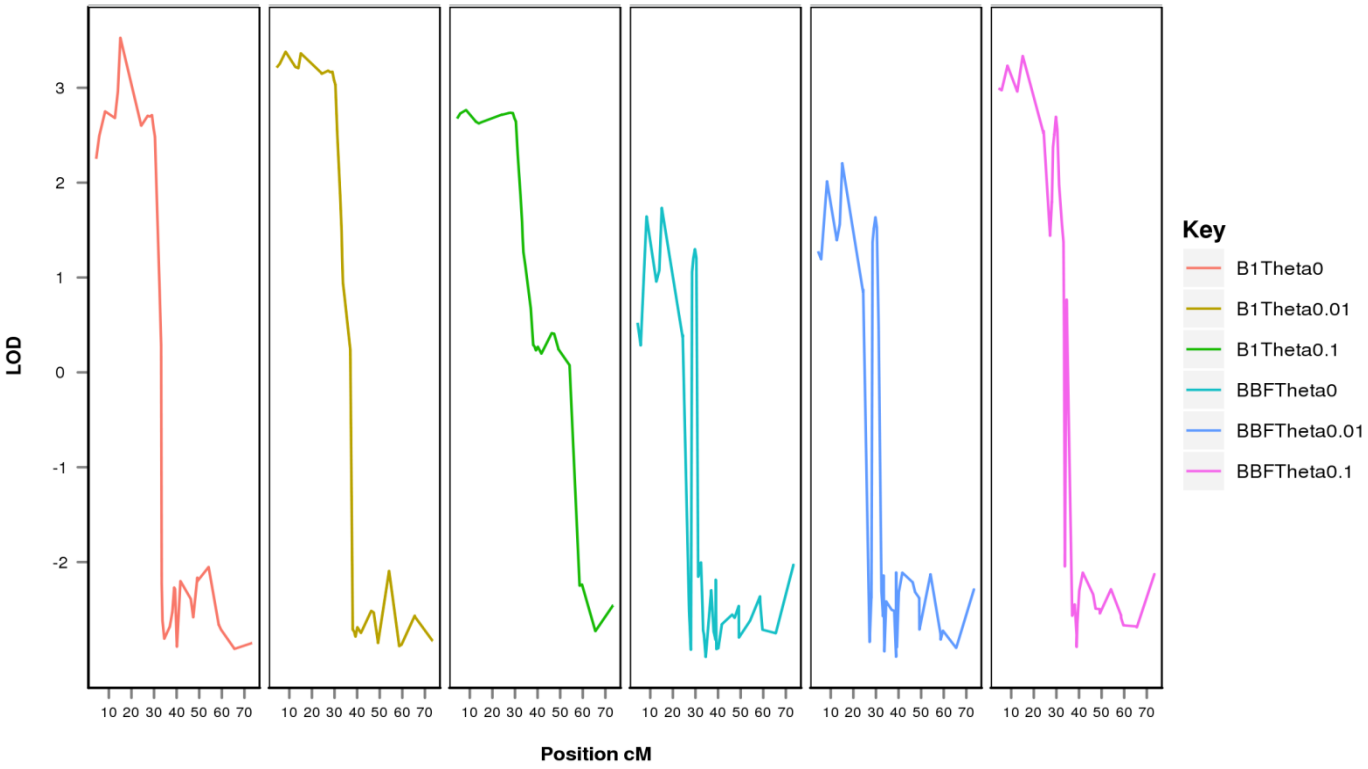


Figure 5.15. McLinkage TLOD scores for Branch 1 (B1) and the BBF at recombination fractions (Theta) ranging from 0.0 through to 0.1 on chromosome 22 versus centiMorgan position. TLOD scores are maximised at recombination fraction of 0.0 in Branch 1 and at 0.1 in the BBF.

5.4.5.2 The Broad Phenotype Model: MERLIN Results

Using the same broad phenotype model and recessive mode of disease transmission, MERLIN analysis of Branch 1 and the BBF identified the same maximum region reported by McLinkage on chromosome 22q11.21-q12.1, where a maximum HLOD score of 3.76 (α=0.76) and 3.37 (α=0.67) was reported for Branch 1 and the BBF respectively, indicating suggestive evidence for linkage. The core region implicated by MERLIN does however differ from that identified by McLinkage, with the former implicating a region spanning 95.4 kb on chromosome 22q12.1 for both Branch 1 and the BBF, approximately 3.8 Mb away from the closest marker in the Mclinkage core region (Figure 5.16, Table 5.6).

Using the broad phenotype model, this region did not report any suggestive or significant HLOD scores under the dominant mode of disease transmission.

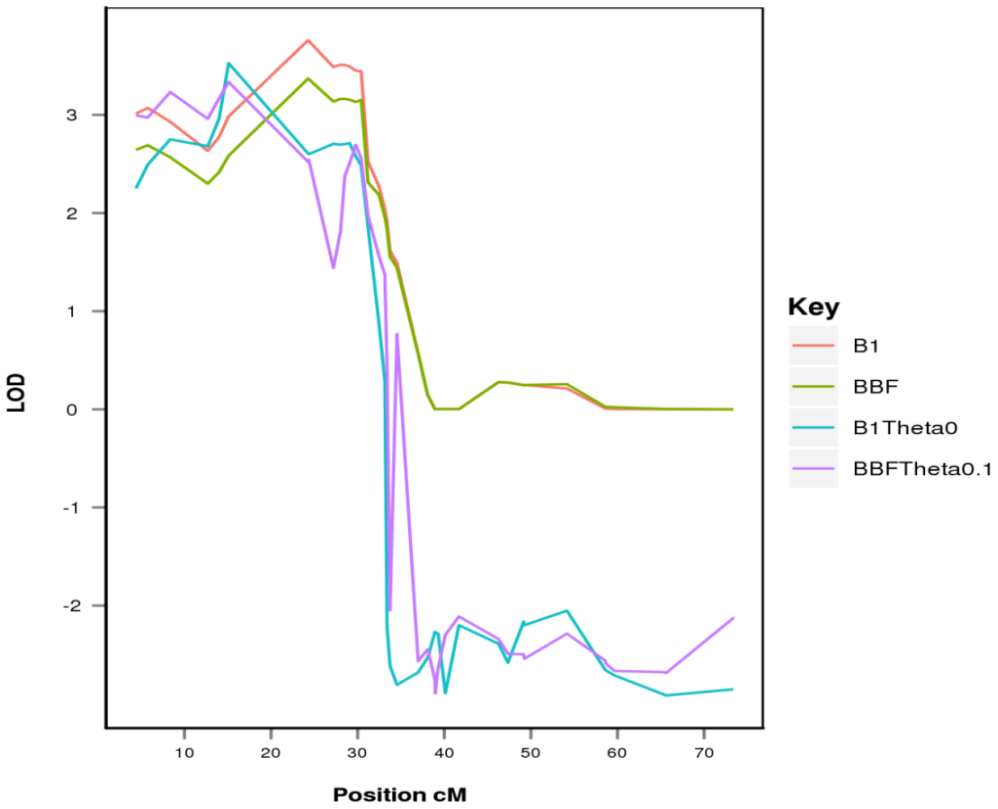


Figure 5.16. MERLIN HLOD scores and McLinkage TLOD scores, at recombination fractions (Theta) of 0.0 for Branch 1 (B1) and 0.1 for the BBF, on chromosome 22 versus centiMorgan position using the broad phenotype model under the recessive mode of disease transmission. The linkage regions identified by the two programs overlap, however the linkage peaks occur approximately 3.8 Mb apart.

SNP	Physical Position	Genetic Position	MERLIN HLOD				McLinkage TLOD	
			Branch 1	α	BBF	α	Branch 1	BBF
rs181402	18229774	4.40	3.01*	0.86	2.64	0.74	2.25	3.00*
rs2401505	18583574	5.77	3.07*	0.84	2.69	0.72	2.49	2.97*
rs2871038	19634134	8.36	2.93*	0.80	2.57	0.69	2.75	3.23*
rs431319	21449028	12.69	2.63	0.75	2.30	0.63	2.68	2.96*
rs448041	21449411	12.70	2.63	0.75	2.30	0.63	2.68	2.96*
rs878825	21982249	14.00	2.78	0.79	2.42	0.65	2.96*	3.17*
rs2014298	22702291	15.10	2.98*	0.84	2.58	0.69	3.53*	3.34*
rs134709	26503944	24.290	3.76*	0.76	3.37*	0.67	2.61	2.52
rs134711	26504037	24.292	3.76*	0.76	3.37*	0.67	2.61	2.52
rs79360	26599293	24.36	3.75*	0.76	3.36*	0.67	2.60	2.54
rs720914	27219881	27.20	3.49*	0.73	3.14*	0.63	2.71	1.44
rs2157465	27381346	27.97	3.51*	0.73	3.16*	0.63	2.70	1.80
rs743958	27388405	28.01	3.51*	0.73	3.16*	0.63	2.70	1.80
rs760531	27493361	28.51	3.51*	0.72	3.16*	0.63	2.70	2.38
rs471216	27610637	29.07	3.49*	0.72	3.16*	0.63	2.71	2.51
rs1401527	27610680	29.07	3.49*	0.72	3.16*	0.63	2.71	2.51
rs2342939	27761128	29.78	3.45*	0.72	3.13*	0.64	2.58	2.69
rs734088	27889627	30.40	3.45*	0.72	3.15*	0.65	2.48	2.56

Table 5.6. MERLIN HLOD and McLinkage TLOD scores at a recombination fraction of zero for Branch 1 and 0.1 for the BBF on chromosome 22q11.21-q12.1 under the broad phenotype model and a recessive mode of disease transmission. One asterisk denotes SNPs with whole genome suggestive evidence for linkage defined as LOD scores greater or equal to 2.90.

5.4.5.2.1 Subfamily Contribution to the Heterogeneity LOD Score

The putative linkage region on chromosome 22q11.21-q12.1 confers susceptibility to disorders defined under the broad phenotype model (BPI, BPII, SAD, BPNOS, cyclothymia) in a proportion of subfamilies from Branch 1 only. Evaluation of subfamily contribution to the overall HLOD score in Branch 1 indicated that only six subfamilies contributed positively to the overall HLOD score. Subfamilies 1, 5, and 7 seemed to be driving the linkage signal in this region by contributing HLOD scores greater than one (nominally significant). Expanding the analysis to the BBF showed that in addition to the contributions of subfamilies from Branch 1, none of the subfamilies in Branches 2 or 3 contributed positively to the linkage signal in this region. These findings are suggestive of heterogeneity at this locus (Figure 5.17) and contrast with the results of the 3p24.3-p24.1 region with subfamilies 1, 5 and 7 having positive HLOD scores for 22q11.21-q12.1 and negative for 3p25.1-p23.2 whereas subfamilies 2 and 3 show the opposite pattern.

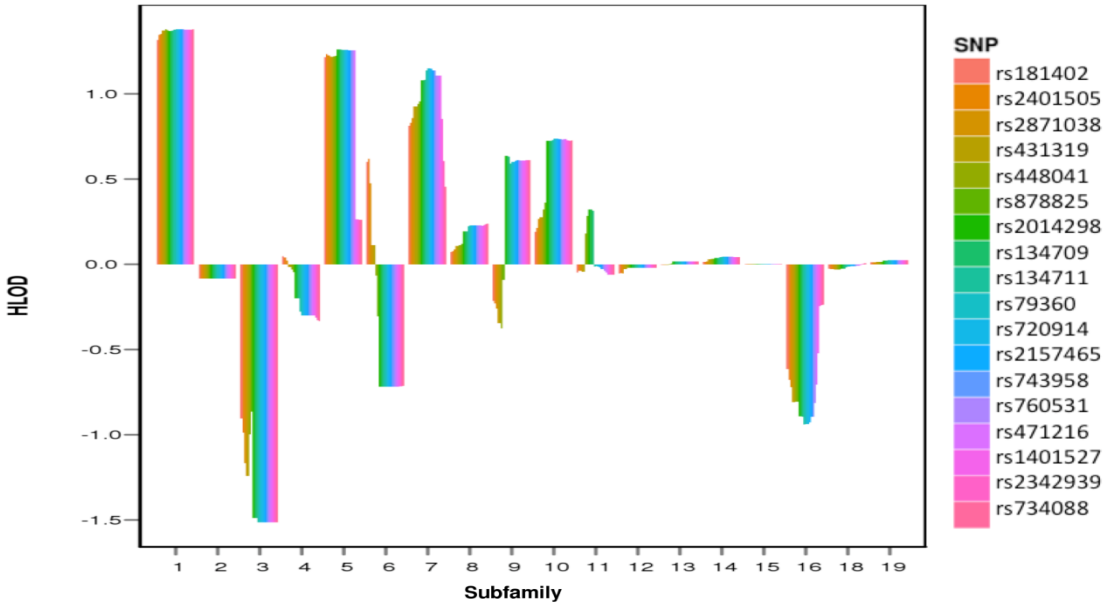


Figure 5.17. Subfamily contribution to the HLOD scores on chromosome 22q11.21-q12.1 under the broad phenotype model. Subfamilies 1 through to 12 constitute Branch 1, subfamilies 13 and 14 constitute Branch 2 and subfamilies 15 through to 19 constitute Branch 3. Subfamily 17 was dropped from the analyses

5.4.5.3 The Broad Phenotype Model: Non-parametric Results

The region on chromosome 22q11.21-q12.1 is the only parametric linkage region supported with some linkage evidence from NPL methods using the same phenotype model. It achieved NPL_{all} and NPL_{pairs} LOD scores greater than one in Branch 1 and the BBF, with a maximum NPL_{all} LOD score of 1.26 (Delta=0.61) and NPL_{pairs} LOD score of 1.0 (Delta=0.95) in Branch 1 and a maximum NPL_{all} LOD score of 1.23 (Delta=0.60) and NPL_{pairs} LOD score of 1.01 (Delta=0.94) in the BBF.

5.4.5.4 The Narrow, Super, and Depression Phenotype Models

The region identified on chromosome 22q11.21-q12.1 using the broad phenotype model yielded no suggestive or significant evidence for linkage using the narrow, super, or depression phenotype models under the dominant or recessive modes of disease transmission in Branch 1 or the BBF.

5.5 Non-Parametric Linkage Analyses

This section will commence with a brief summary of the statistics employed by MERLIN in the analyses with the aim of clarifying their use in the study. Chromosomal regions achieving the corrected LOD scores for significant (LOD scores greater or equal to 3.7) and suggestive (LOD scores greater or equal to 2.7) linkage will then be presented per chromosome.

5.5.1 Non-parametric Linkage Statistics Linkage Statistics

Nonparametric linkage using MERLIN performs an NPL_{all} and an NPL_{pairs} test. NPL_{all} estimates Identity-by-Decent (IBD) allele sharing across groups of affected family members and therefore only attributes significance to regions where a single variant confers susceptibility. Conversely NPL_{pairs} estimates IBD between pairs of affected family members and therefore attributes significance to regions where multiple variants infer susceptibility in different pairs of affected family members. In other words, NPL_{pairs} allows for allelic heterogeneity within family members in each subfamily, while NPL_{all} does not. LOD scores from both statistics will be presented in this section. The allele sharing parameter Delta (Kong & Cox, 1997), which measures excess IBD allele sharing between pairs and groups of affected family members under NPL_{pairs} and NPL_{all} respectively will also be reported here. As a rule, Delta scores above zero indicate excess in IBD sharing. We will now explore significant and suggestive non-parametric linkage results per chromosome.

5.5.2 Chromosome 1p

5.5.2.1 The Narrow Phenotype Model

NPL analysis undertaken on Branch 1 yielded a maximum NPL_{all} LOD score of 2.96 supportive of whole genome suggestive linkage in a core region of 9.4 Mb on chromosome 1p22.2-p21.3 using the narrow phenotype model (Figure 5.18). This was reported using the NPL_{all} statistic only, indicating excess IBD allele sharing between affected groups in Branch 1 (Delta=2.12). Expanding the region to include markers with a 1-LOD drop in threshold from 2.96 identified a maximum putative susceptibility locus of 10.7 Mb on chromosome 1p22.2-p21.2. The linkage signal was specific to Branch 1 as analysis of the BBF substantially decreased the NPL_{all} LOD score to a maximum of 2.12 (Delta=0.98), which does not meet criteria for suggestive linkage (Table 5.7).

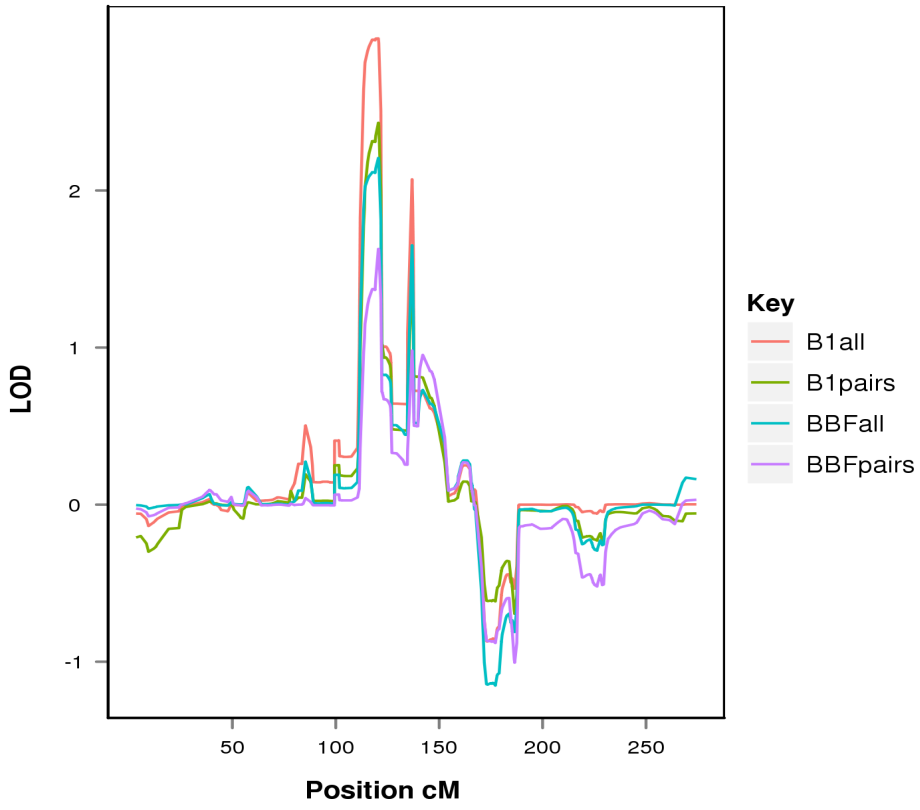


Figure 5.18. NPL_{all} and NPL_{pairs} exponential LOD scores for Branch 1 (B1) and the BBF on chromosome 1 versus centiMorgan position under the narrow phenotype model.

SNP	Physical Position	Genetic Position	Branch 1				BBF			
			NPL_{all}	Delta	NPL_{pairs}	Delta	NPL_{all}	Delta	NPL_{pairs}	Delta
rs1074541	89509499	113.46	2.64	1.47	1.78	1.47	1.87	0.95	0.97	1.03
rs1380416	89949213	113.88	2.75*	1.49	1.89	1.49	1.97	0.95	1.06	1.06
rs2390795	90139260	114.06	2.80 *	1.52	1.99	1.53	2.01	0.96	1.13	1.09
rs519989	90223400	114.14	2.82*	1.53	2.02	1.54	2.02	0.96	1.15	1.09
rs716988	91543728	115.42	2.89*	1.59	2.18	1.60	2.07	0.97	1.27	1.14
rs1926261	92162920	116.09	2.92*	1.61	2.24	1.62	2.09	0.98	1.31	1.15
rs546550	94550555	117.62	2.96*	1.64	2.31	1.63	2.12	0.98	1.37	1.16
rs545542	94550696	117.62	2.96*	1.64	2.31	1.63	2.12	0.98	1.37	1.16
rs952499	94558425	117.62	2.96*	1.64	2.31	1.63	2.12	0.98	1.37	1.16
rs951534	94625274	117.68	2.96*	1.64	2.31	1.63	2.12	0.98	1.37	1.16
rs1411698	94686110	117.73	2.96*	1.64	2.31	1.63	2.12	0.98	1.37	1.16
rs1073364	94828209	117.84	2.96 *	1.64	2.31	1.63	2.12	0.98	1.37	1.16
rs696619	95005220	117.98	2.96 *	1.64	2.31	1.63	2.12	0.98	1.37	1.16
rs2040074	95158357	118.11	2.96 *	1.64	2.31	1.63	2.12	0.98	1.37	1.16
rs1414904	95633977	118.50	2.96 *	1.64	2.31	1.63	2.12	0.98	1.37	1.16
rs1113271	96168570	118.88	2.96 *	1.64	2.31	1.63	2.11	0.98	1.37	1.16
rs953401	96186671	118.89	2.96 *	1.64	2.31	1.63	2.11	0.98	1.37	1.16
rs1935571	96414335	118.98	2.96 *	1.64	2.31	1.63	2.11	0.98	1.37	1.16
rs321241	96476493	119.01	2.96 *	1.64	2.31	1.63	2.11	0.98	1.37	1.16
rs1144305	96506730	119.02	2.96 *	1.64	2.31	1.63	2.11	0.98	1.37	1.16
rs54790	97265741	119.34	2.96 *	1.67	2.35	1.66	2.14	0.99	1.44	1.20
rs728656	98855166	120.58	2.97 *	1.78	2.43	1.74	2.21	1.04	1.63	1.30
rs764051	99356958	120.80	2.94 *	1.80	2.41	1.75	2.19	1.04	1.62	1.31
rs951125	100198059	121.81	2.50	1.80	1.99	1.72	1.81	1.05	1.31	1.28

Table 5.7. NPL_{all} and NPL_{pairs} exponential LOD scores and allele sharing score (Delta) reported for Branch 1 and the BBF on chromosome 1p22.2-p21.2 using the narrow phenotype model. One asterisk denotes SNPs with whole-genome suggestive evidence for linkage defined as LOD scores greater or equal to 2.7.

5.5.2.1.1 Subfamily Contribution to the LOD Score

The putative linkage region on chromosome 1p22.2-p21.2 confers susceptibility to disorders defined under the narrow phenotype model (BPI, BPII, and SAD) in a proportion of subfamilies in Branch 1 only. Evaluation of subfamily contribution to the overall LOD score indicated that only seven subfamilies from Branch 1 contributed positively to the overall LOD score, with only subfamily 1 achieving nominally significant scores (i.e. LOD scores greater or equal to 0.59 equivalent to p-value of 0.05). Further, expanding the analysis to the BBF showed that in addition to the contributions of subfamilies from Branch 1 none of the subfamilies in Branches 2 or 3 contributed positively to the linkage signal in this region (Figure 5.19).

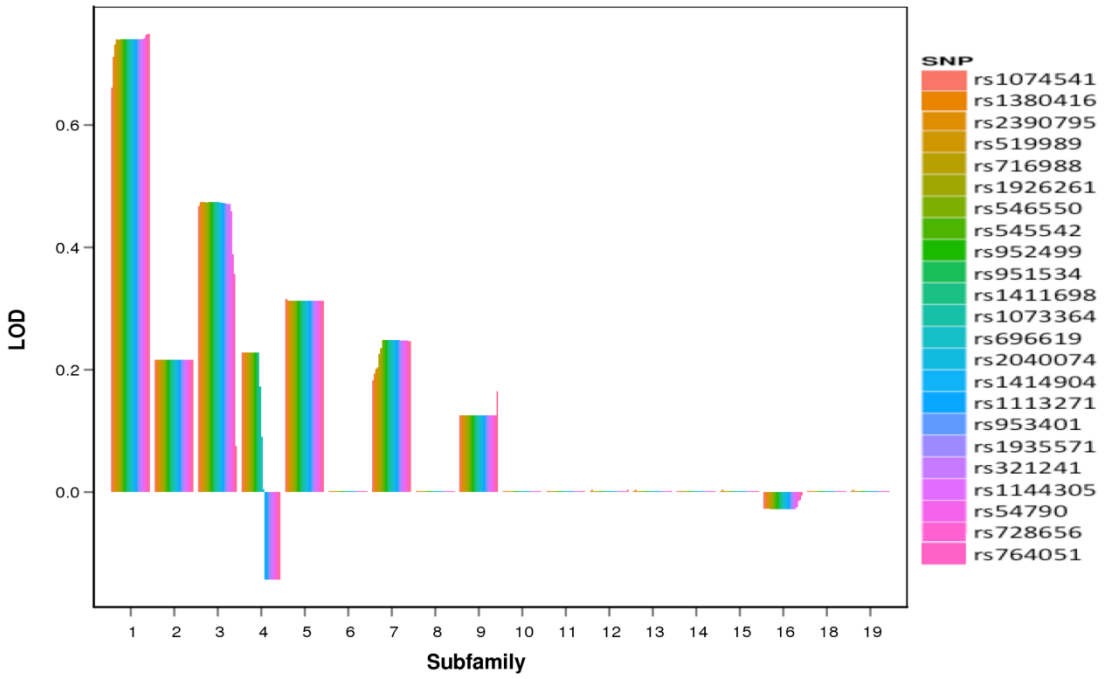


Figure 5.19. Subfamily contribution to the NPL_{all} LOD score on chromosome 1p22.2-p21.2 using the narrow phenotype model. Subfamilies 1 through to 12 constitute Branch 1, subfamilies 13 and 14 constitute Branch 2 and subfamilies 15 through to 19 constitute Branch 3. Subfamily 17 was dropped from the analyses.

5.5.2.2 The Broad Phenotype Model

A subset of SNPs in the core region identified using the narrow phenotype model showed suggestive evidence for linkage using the broad phenotype model in Branch 1 (Figure 5.20). Using this phenotype model refined the core linkage peak to a narrower region of approximately 5.6 Mb on chromosome 1p22.2-p21.3 with a maximum NPL_{all} LOD score of 2.93 and a maximum NPL_{pairs} LOD score of 2.70, indicating excess allele sharing between groups (Delta=0.94) and pairs (Delta =1.49) of affected family members respectively. Expanding the region to include markers with a 1-LOD drop in threshold identified a maximum susceptibility locus of 9.8 Mb on chromosome 1p22.2-p21.3, 900 kb smaller than the maximum region defined under the narrow phenotype model (Tables 5.8).

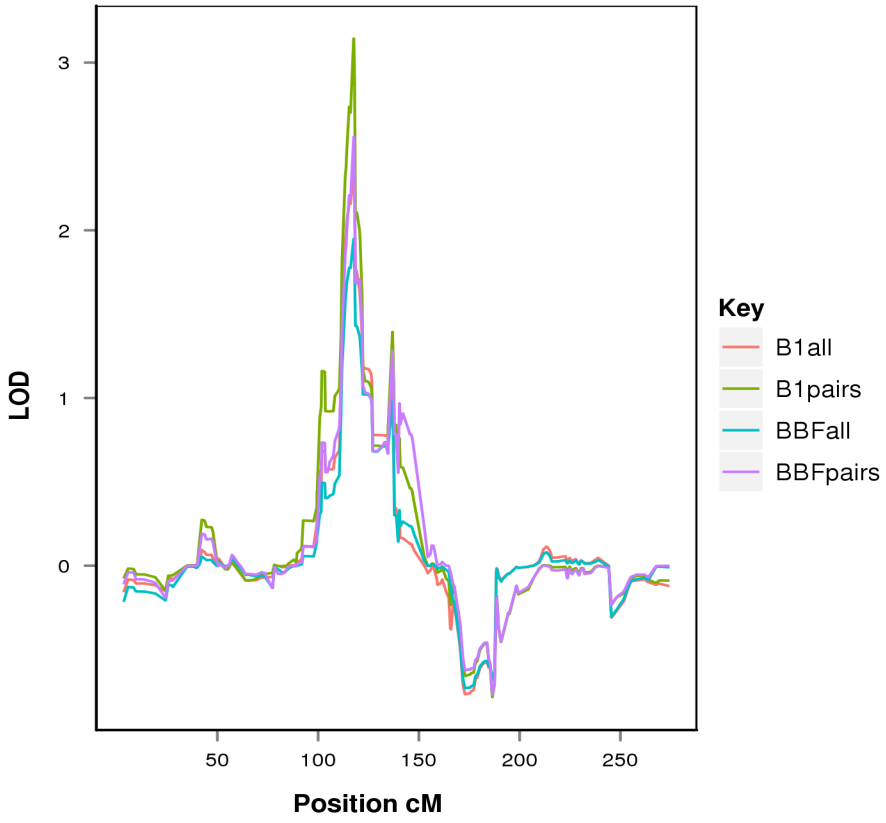


Figure 5.20. NPL_{all} and NPL_{pairs} exponential LOD scores for Branch 1 (B1) and the BBF on chromosome 1 versus centiMorgan position under the broad phenotype model.

SNP	Physical Position	cM Position	Branch 1				BBF			
			NPL_{all}	Delta	NPL_{pairs}	Delta	NPL_{all}	Delta	NPL_{pairs}	Delta
rs1074541	89509499	113.46	2.75*	0.92	2.33	1.42	2.12	0.71	1.69	1.12
rs1380416	89949213	113.88	2.75*	0.93	2.42	1.43	2.20	0.71	1.77	1.13
rs2390795	90139260	114.06	2.86*	0.93	2.48	1.44	2.23	0.71	1.82	1.14
rs519989	90223400	114.14	2.88*	0.93	2.50	1.44	2.24	0.71	1.83	1.14
rs716988	91543728	115.42	2.90*	0.93	2.60	1.47	2.26	0.71	1.92	1.16
rs1926261	92162920	116.09	2.91*	0.94	2.64	1.48	2.27	0.71	1.95	1.17
rs546550	94550555	117.62	2.93*	0.94	2.70*	1.49	2.28	0.71	1.99	1.18
rs545542	94550696	117.62	2.93*	0.94	2.70*	1.49	2.28	0.71	1.99	1.18
rs952499	94558425	117.62	2.93*	0.94	2.70*	1.49	2.28	0.71	1.99	1.18
rs951534	94625274	117.68	2.93*	0.94	2.70*	1.49	2.28	0.71	1.99	1.18
rs1411698	94686110	117.73	2.93*	0.94	2.70*	1.49	2.28	0.71	1.99	1.18
rs1073364	94828209	117.84	2.88*	0.93	2.63	1.49	2.23	0.71	1.93	1.17
rs696619	95005220	117.98	2.80*	0.92	2.54	1.48	2.17	0.71	1.85	1.16
rs2040074	95158357	118.11	2.72*	0.91	2.43	1.47	2.10	0.70	1.76	1.15
rs1414904	95633977	118.50	2.30	0.81	1.71	1.19	1.80	0.66	1.23	0.95
rs1113271	96168570	118.88	2.29	0.81	1.70	1.19	1.80	0.66	1.23	0.95
rs953401	96186671	118.89	2.29	0.81	1.70	1.19	1.80	0.66	1.23	0.95
rs1935571	96414335	118.98	2.29	0.81	1.70	1.19	1.79	0.66	1.22	0.95
rs321241	96476493	119.01	2.29	0.81	1.70	1.19	1.79	0.66	1.22	0.95
rs1144305	96506730	119.02	2.29	0.81	1.70	1.19	1.79	0.66	1.22	0.95
rs54790	97265741	119.34	2.29	0.81	1.72	1.20	1.80	0.66	1.26	0.97
rs728656	98855166	120.58	2.25	0.82	1.73	1.23	1.80	0.67	1.37	1.02
rs764051	99356958	120.80	2.21	0.82	1.71	1.23	1.78	0.67	1.36	1.02
rs951125	100198059	121.81	1.86	0.80	1.53	1.25	1.48	0.65	1.24	1.03

Table 5.8. NPL_{all} and NPL_{pairs} exponential LOD scores and allele sharing scores (Delta) for chromosome 1p22.2-p21.3 using the broad phenotype model for Branch 1 and BBF. One asterisk denotes SNPs with whole-genome suggestive evidence for linkage defined as LOD scores greater or equal to 2.7.

5.5.2.2.1 Subfamily Contribution to the LOD Score

The putative region on chromosome 1p22.2-p21.3 confers susceptibility to disorders defined under the broad phenotype model (BPI, BPII, SAD, BPNOS, cyclothymia) in a proportion of subfamilies from Branch 1 only. Evaluation of subfamily contribution to the overall LOD score indicated that only eight subfamilies from Branch 1 contributed positively to the overall LOD score. Similar to findings from the narrow phenotype model, only subfamily 1 achieved nominally significant scores (i.e. LOD scores greater or equal to 0.59 equivalent to p-value of 0.05) in this region. Expanding the analysis to the BBF showed that in addition to the contributions of subfamilies from Branch 1, none of the subfamilies from Branches 2 or 3 contributed positively to the LOD score (Figure 5.21).

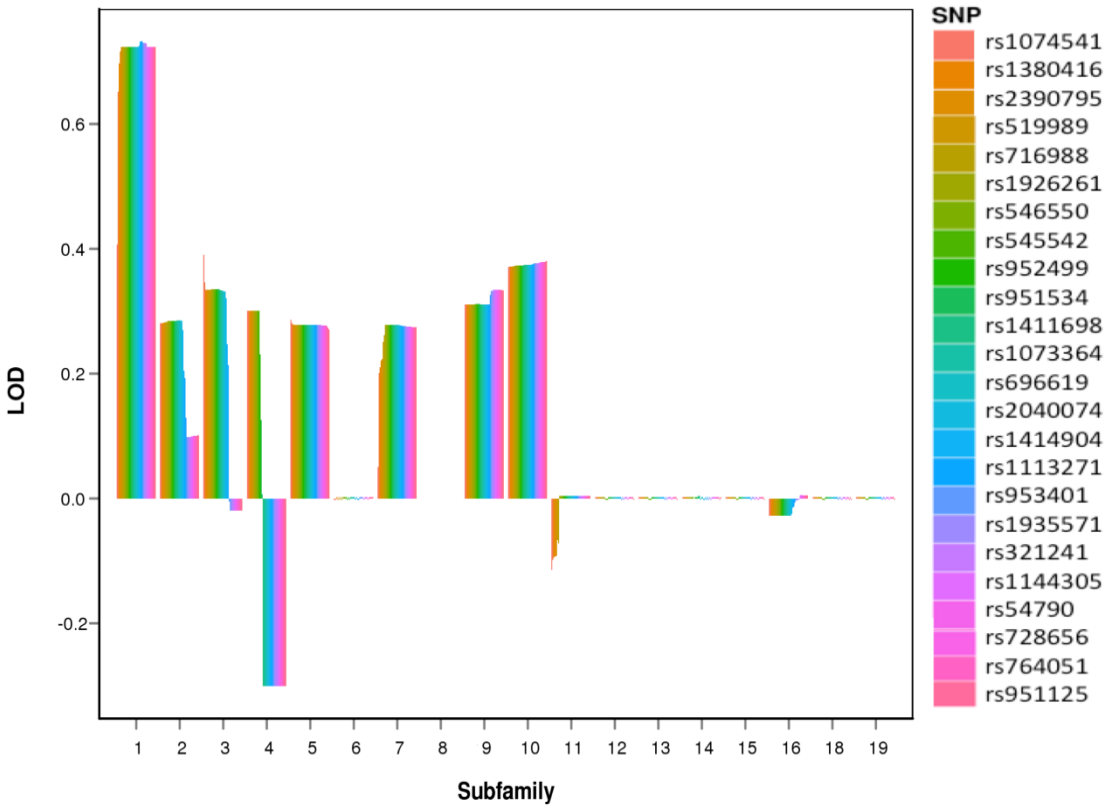


Figure 5.21. Subfamily contribution to the NPL_{all} LOD score on chromosome 1p22.2-p22.3 under the broad phenotype model. Subfamilies 1 through to 12 constitute Branch 1, subfamilies 13 and 14 constitute Branch 2 and subfamilies 15 through to 19 constitute Branch 3. Subfamily 17 was dropped from the analyses.

5.5.2.3 The Super and Depression Phenotype Models

The linkage region identified on chromosome 1p22.2-p21.3 using the narrow and broad phenotype models yielded no suggestive or significant evidence for linkage using the super or depression phenotype models in Branch 1 or the BBF. The only positive LOD scores greater or equal to 1 were achieved in analysis under the super phenotype model, which yielded a maximum NPL_{pairs} LOD score of 1.19. These results implicate chromosome 1p22.2-p21.3 as a putative susceptibility locus for disorders defined under the broad phenotype model (BPI, BPI, SAD, BPNOS cyclothymia).

5.5.3 Chromosome 1q

5.5.3.1 The Super Phenotype Model

A region on chromosome 1q21.1-q21.2, spanning 1.1 Mb, was identified with suggestive evidence for linkage in the BBF with a maximum NPL_{pairs} LOD score of 2.83 under the super phenotype model (Figure 5.22). These results were achieved using the NPL_{pairs} statistic only, indicating increased allele sharing among pairs of affected family members only (Delta=0.95). A maximum region including SNPs with a 1-LOD drop in threshold from the maximum LOD of 2.83 was identified on chromosome 1q21.1-q21.3 and expanded the putative region to approximately 8 Mb. This region was not corroborated in Branch 1, yielding less than whole genome suggestive LOD scores (Table 5.9). These results implicate chromosome 1q21.1-q21.3 as a putative susceptibility locus for mood disorders (BPI, BPII, SAD, BPNOS cyclothymia) in the BBF.

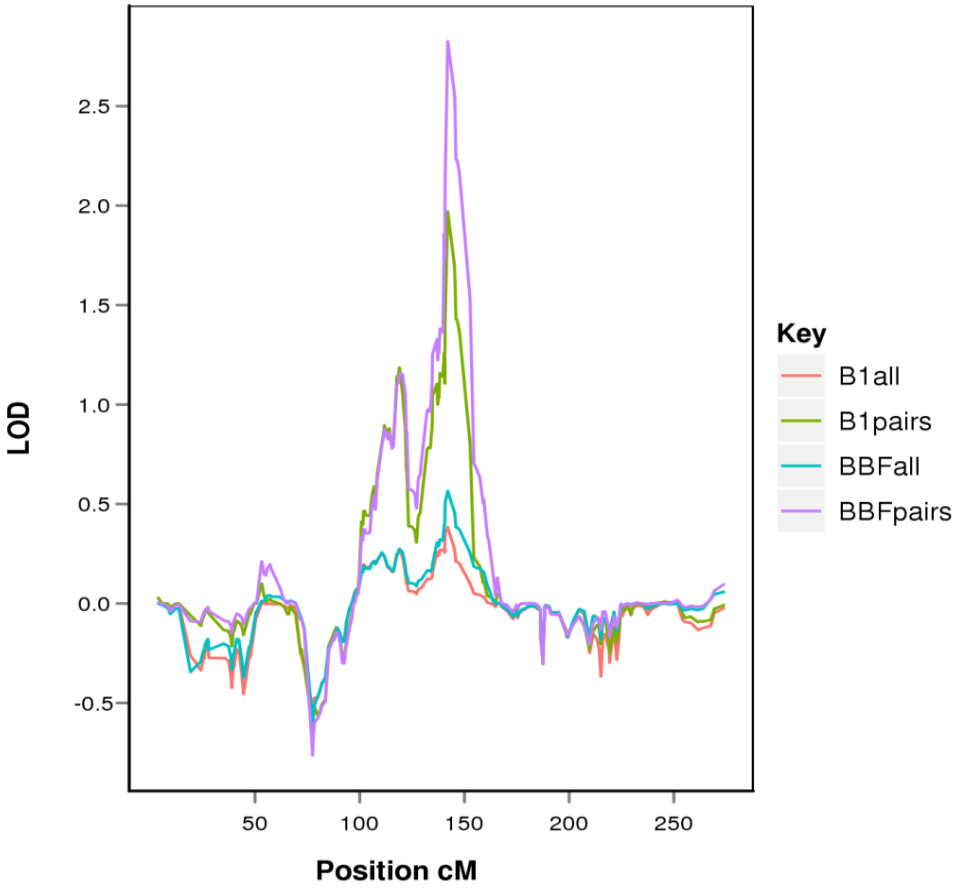


Figure 5.22. NPL_{all} and NPL_{pairs} exponential LOD scores for Branch 1 (B1) and the BBF on chromosome 1 versus centiMorgan position under the super phenotype model.

SNP	Physical Position	cM Position	Branch 1				BBF			
			NPL_{all}	Delta	NPL_{pairs}	Delta	NPL_{all}	Delta	NPL_{pairs}	Delta
rs719969	146555660	140.78	0.36	0.18	1.43	0.76	0.51	0.19	2.13	0.85
rs2077749	146652637	142.04	0.38	0.18	1.97	0.87	0.57	0.19	2.83*	0.95
rs720609	146653028	142.05	0.38	0.18	1.97	0.87	0.57	0.19	2.83 *	0.95
rs953087	146739540	142.05	0.38	0.18	1.97	0.87	0.57	0.19	2.83 *	0.95
Rs1832116	147779090	142.09	0.38	0.18	1.97	0.87	0.56	0.19	2.82 *	0.95
Rs1196462	151627604	145.24	0.27	0.17	1.70	0.82	0.46	0.18	2.56	0.91
Rs2338019	151956814	145.53	0.25	0.17	1.62	0.80	0.44	0.18	2.48	0.90
rs726865	152400055	145.94	0.21	0.16	1.43	0.74	0.38	0.17	2.23	0.84
rs726863	152400361	145.94	0.21	0.16	1.43	0.74	0.38	0.17	2.23	0.84
rs950338	152756806	146.26	0.21	0.16	1.43	0.74	0.38	0.17	2.23	0.84
rs1560832	153335384	146.78	0.21	0.16	1.43	0.74	0.38	0.17	2.22	0.84
rs724781	153336018	146.79	0.21	0.16	1.43	0.74	0.38	0.17	2.22	0.84
Rs2335230	154589628	147.67	0.20	0.15	1.36	0.73	0.37	0.17	2.16	0.84

Table 5.9. NPL_{all} and NPL_{pairs} exponential LOD scores and allele sharing scores (Delta) for chromosome 1q21.1-q21.3 under the super phenotype model for Branch 1 and the BBF. One asterisk denotes SNPs with whole-genome suggestive evidence for linkage defined as LOD scores greater or equal to 2.7.

5.5.3.1.1 Subfamily Contribution to the LOD score

The putative linkage region on chromosome 1q21.1-q21.3 confers susceptibility to disorders defined under the super phenotype model (BPI, BPII, SAD, BPNOS, cyclothymia, MDD, dysthymia) in a proportion of subfamilies from Branch 1 and the BBF. Evaluation of subfamily contribution to the overall LOD score indicated that ten subfamilies from Branch 1 contributed positively, albeit in different proportions, to the overall LOD score in the analyses conducted in Branch 1. Further, expanding the analysis to the BBF showed that, in addition to the contributions of subfamilies from Branch 1, two subfamilies from Branches 2 (subfamily 14) and 3 (subfamily16) contributed positively to the overall LOD score. However, none of the subfamilies reported nominally significant LOD scores in the region (Figure 5.23).

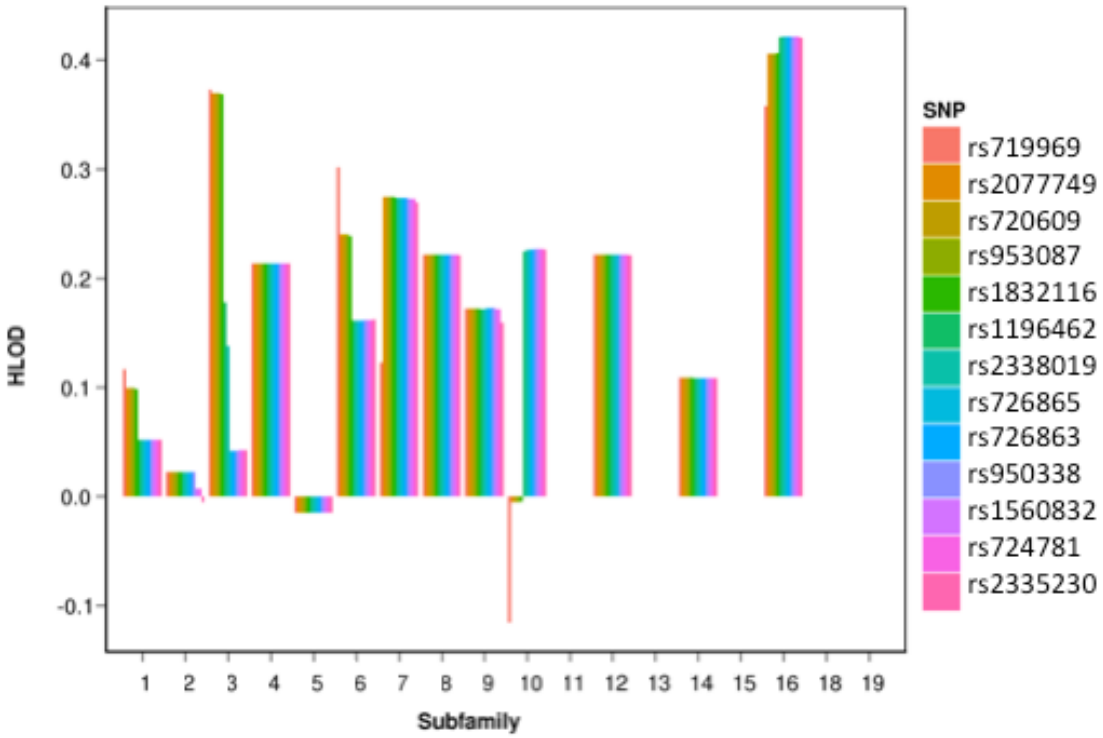


Figure 5.23. Subfamily contribution to the NPL_{pairs} LOD score on chromosome 1q21.1-q21.3. Subfamilies 1 through to 12 constitute Branch 1, subfamilies 13 and 14 constitute Branch 2 and subfamilies 15 through to 19 constitute Branch 3. Subfamily 17 was dropped from the analyses.

5.5.3.2 The Narrow, Broad, and Depression Phenotype Models

The linkage region identified on chromosome 1q21.1-q21.3 using the super phenotype model yielded LOD scores that did not meet criteria for suggestive or significant linkage using the narrow, broad, or depression phenotype models.

5.5.4 Chromosome 2p

5.5.4.1 The Super Phenotype Model

A region spanning 3 Mb on chromosome 2p23.1-p22.3 achieved whole genome significant linkage under the super phenotype model (Figure 5.24). NPL_{pairs} produced a maximum LOD score of 3.83 in Branch 1, while NPL_{all} yielded non-significant LOD scores at this locus, indicating that increased allele sharing was only present between pairs of affected family members (Delta=1.10). A maximum region defined by a 1-LOD drop in threshold from the maximum LOD of 3.83 identified a larger region on chromosome 2p23.1-p22.3 spanning 5.4 Mb. Analysis of the BBF decreased the LOD scores to levels indicative of suggestive linkage only, with a maximum NPL_{pairs} score of 3.46 (Table 5.10). These results implicate chromosome 2p23.1-p22.3 as a putative susceptibility locus for mood disorders identified under the super phenotype model (BPI, BPII, SAD, BPNOS cyclothymia, MDD, and dysthymia).

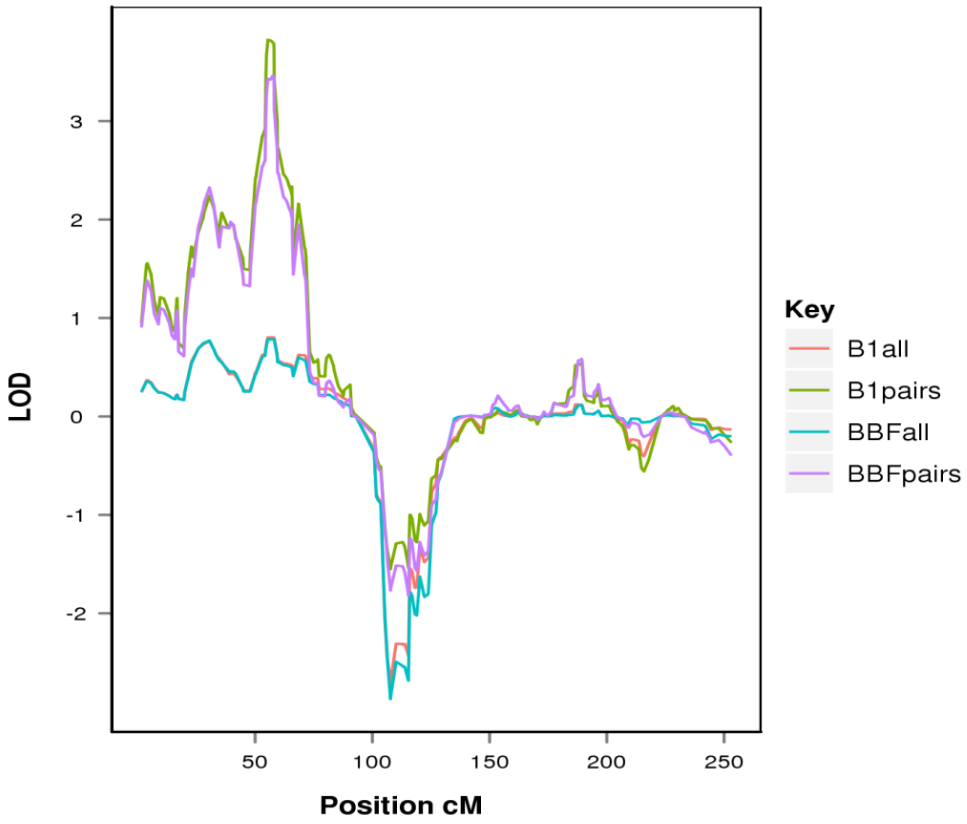


Figure 5.24. NPL_{all} and NPL_{pairs} exponential LOD scores on chromosome 2 versus centiMorgan (cM) position under the super phenotype model

SNP	Physical Position	cM Position	Branch1				BBF			
			NPL_{all}	Delta	NPL_{pairs}	Delta	NPL_{all}	Delta	NPL_{pairs}	Delta
rs1509577	30410818	53.30	0.63	0.20	2.85	0.99	0.61	0.61	2.55	0.90
rs829669	30720752	53.82	0.63	0.20	2.89	0.99	0.61	0.61	2.58	0.90
rs499476	30985774	54.26	0.64	0.20	2.92*	0.99	0.62	0.62	2.60	0.91
rs2364445	31176516	54.58	0.70	0.20	3.43*	1.08	0.68	0.68	3.06	0.98
rs1862975	31326432	54.84	0.74	0.20	3.67*	1.11	0.72	0.72	3.27	1.00
rs2365842	31588288	55.34	0.80	0.20	3.80**	1.10	0.78	0.78	3.40*	1.00
rs206847	31611367	55.37	0.80	0.20	3.81**	1.10	0.78	0.78	3.41*	1.00
rs2208532	31788989	55.44	0.80	0.20	3.83**	1.10	0.78	0.78	3.43*	1.00
rs564310	31892342	55.48	0.80	0.20	3.83**	1.10	0.78	0.78	3.43*	1.00
rs219085	33292452	56.02	0.80	0.20	3.82**	1.10	0.78	0.78	3.42*	1.00
rs723042	33541702	56.34	0.80	0.20	3.82**	1.10	0.78	0.78	3.42*	1.00
rs723043	33541787	56.34	0.80	0.20	3.82**	1.10	0.78	0.78	3.42*	1.00
rs602622	33658226	56.54	0.80	0.20	3.82**	1.10	0.78	0.78	3.42*	1.00
rs2197804	33916066	56.98	0.80	0.20	3.81**	1.10	0.78	0.78	3.44*	1.01
rs1961268	34249225	57.56	0.80	0.20	3.80**	1.10	0.78	0.78	3.46*	1.01
rs722289	34265233	57.59	0.80	0.20	3.80**	1.10	0.78	0.78	3.46*	1.01
rs2371923	34474473	57.93	0.80	0.20	3.79**	1.10	0.78	0.78	3.45*	1.01
rs3845765	34531775	58.00	0.80	0.20	3.78**	1.10	0.78	0.78	3.45*	1.01
rs3910275	34541186	58.01	0.80	0.20	3.78**	1.10	0.78	0.78	3.44*	1.01
rs3845767	34543150	58.01	0.80	0.20	3.78**	1.10	0.78	0.78	3.44*	1.01
rs1583366	34601927	58.08	0.80	0.20	3.74**	1.10	0.78	0.78	3.40*	1.00
rs892457	34667721	58.15	0.79	0.20	3.43*	1.06	0.78	0.78	3.12*	0.97
rs1371422	35772759	59.44	0.63	0.19	3.00*	1.02	0.61	0.61	2.72*	0.93
rs1371427	35779358	59.45	0.63	0.19	3.00*	1.02	0.61	0.61	2.71*	0.93
rs1371426	35779431	59.46	0.63	0.19	3.00*	1.02	0.61	0.61	2.71*	0.93

Table 5.10. NPL_{all} and NPL_{pairs} exponential LOD scores and allele sharing scores (Delta) on chromosome 2p23.1-p22.3 under the super phenotype model. Two asterisks denote SNPs with whole-genome significance defined with LOD scores greater or equal to 3.7 and one asterisk denotes SNPs with whole-genome suggestive defined with LOD scores greater or equal to 2.7.

5.5.4.1.1 Subfamily Contribution to the LOD Score

The significant linkage region identified on chromosome 2p23.1-p22.3 confers susceptibility to disorders defined under the super phenotype model (BPI, BPII, SAD, BPNOS, cyclothymia, MDD, dysthymia) in a proportion of subfamilies from Branch 1 and the BBF. Evaluation of subfamily contribution to the overall LOD score indicated that ten subfamilies from Branch 1 contributed positively, albeit in different proportions, to the overall LOD score signal in analyses conducted in Branch 1. Expanding the analysis to the BBF showed that in addition to the contributions of subfamilies from Branch 1 only subfamily 14 from Branch 2 contributed positively to the overall LOD score (Figure 5.25). However, none of the subfamilies reported nominally significant linkage to the region defined by LOD scores greater or equal to 0.59.

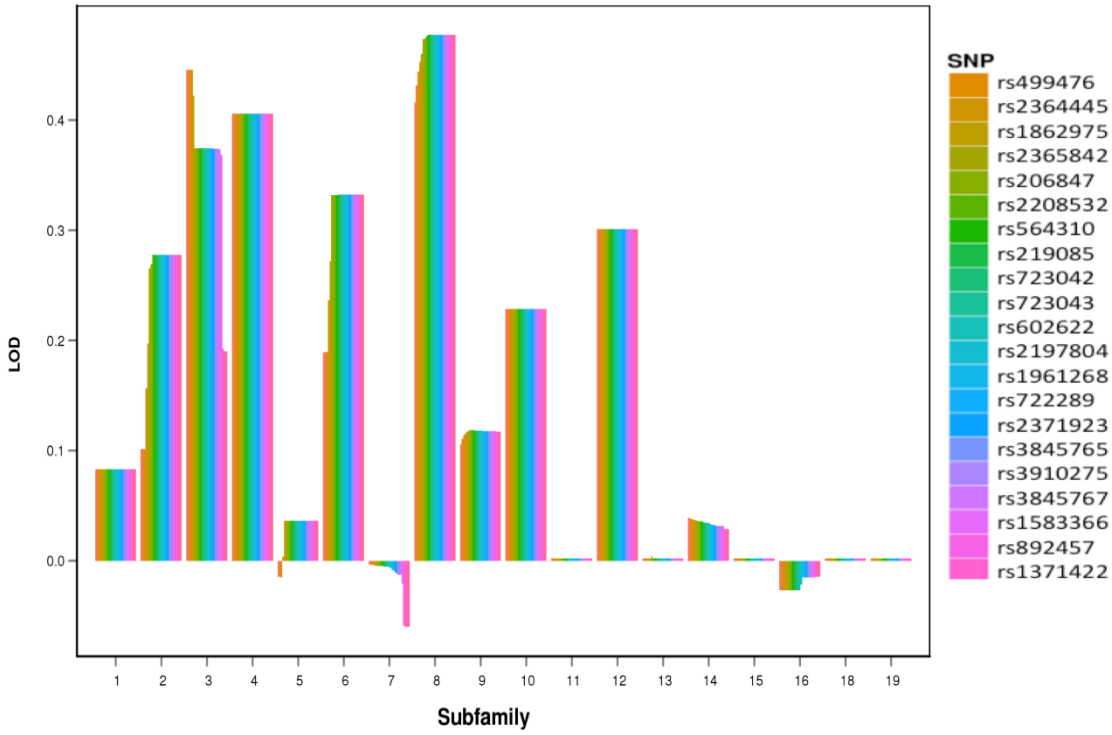


Figure 5.25. Subfamily contribution to the NPL_{pairs} LOD scores on chromosome 2p23.1-p22. Subfamilies 1 through to 12 constitute Branch 1, subfamilies 13 and 14 constitute Branch 2 and subfamilies 15 through to 19 constitute Branch 3. Subfamily 17 was dropped from the analyses.

5.5.4.2 The Narrow, Broad, and Depression Phenotype Models

The region identified on chromosome 2p23.1-p22 using the super phenotype model yielded no suggestive or significant evidence for linkage using the narrow, broad, or depression phenotype models. Analysis under the depression phenotype model, however, yielded a maximum NPL_{pairs} LOD score of 1.18, which only achieves nominal significance.

5.5.5 Chromosome 11p

5.5.5.1 The Depression Phenotype Model

The second highest LOD scores achieved in the entire linkage study occurred in a telomeric region spanning 73.8 kb on chromosome 11p15.4 using the depression phenotype model. NPL analysis revealed whole genome significant linkage using both NPL_{all} and NPL_{pairs} test statistics, indicating excess IBD allele sharing among groups and pairs of affected family members in Branch 1 and the BBF. A maximum LOD score of 4.49 (Delta=1.28) in Branch 1 and 4.70 (Delta=1.33) in the BBF was achieved using the NPL_{all} test. This region was expanded to a maximum region of 2.4 Mb when a 1-LOD drop in threshold was considered (Table 5.11, Figure 5.26).

SNP	Physical Position	cM Position	Branch1				BBF			
			NPL_{all}	Delta	NPL_{pairs}	Delta	NPL_{all}	Delta	NPL_{pairs}	Delta
rs1822285	4232627	6.18	3.08*	1.41	2.87*	2.28	3.30*	1.48	3.11*	2.31
rs968856	5260576	7.98	1.17	0.53	1.85	1.13	1.32	0.54	2.07	1.15
rs956471	5470716	8.40	4.39**	1.29	4.03**	2.00	4.61**	1.33	4.27**	2.02
rs979752	5539139	8.54	4.49**	1.28	4.13**	1.98	4.70**	1.33	4.36**	2.00
rs1433567	5544512	8.55	4.49**	1.28	4.12**	1.98	4.70**	1.33	4.36**	2.00
rs2345344	6582512	10.46	3.41*	1.02	2.53	1.38	3.60*	1.05	2.74*	1.39
rs1354801	6727823	10.78	2.96*	0.95	2.26	1.27	3.15*	0.99	2.47	1.29

Table 5.11. NPL_{all} and NPL_{pairs} exponential LOD scores and allele sharing scores (Delta) on chromosome 11p15.4. Two asterisks denote SNPs reaching whole-genome significance defined with LOD scores greater or equal to 3.7 and one asterisk denotes SNPs with whole-genome suggestive defined with LOD scores greater or equal to 2.7.

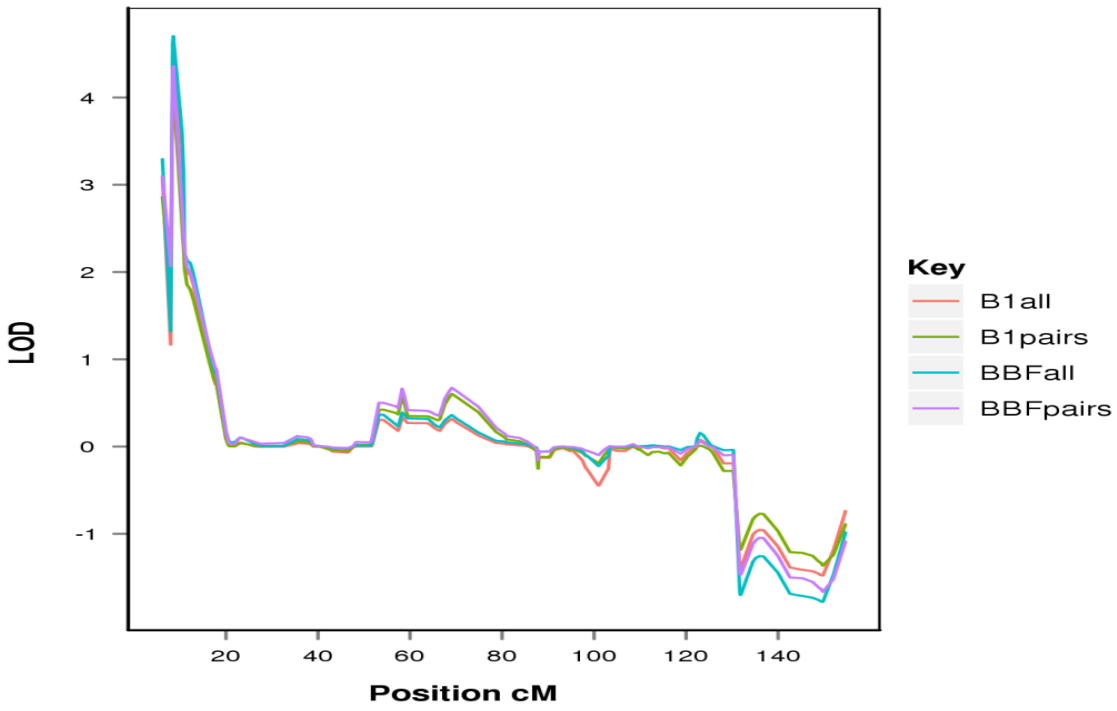


Figure 5.26. NPL_{all} and NPL_{pairs} exponential LOD scores for Branch 1 (B1) and the BBF on chromosome 11 versus centiMorgan position under the depression phenotype model.

5.5.5.1.1 Subfamily Contribution to the LOD Score

The significant linkage region on chromosome 11p15.4 confers susceptibility to disorders defined under the depression phenotype model mainly in subfamilies from Branch 1, as examining subfamily contribution to the overall LOD score revealed that eight subfamilies contributed positively to the overall LOD score, with subfamily 7 achieving nominal significance (LOD greater or equal to 0.59) in the region. Expanding the analysis to the BBF showed that in addition to the contributions of subfamilies from Branch 1 only subfamily 16 from Branch 3 contributed positively to the overall LOD score (Figure 5.27). These findings are suggestive of heterogeneity at this locus.

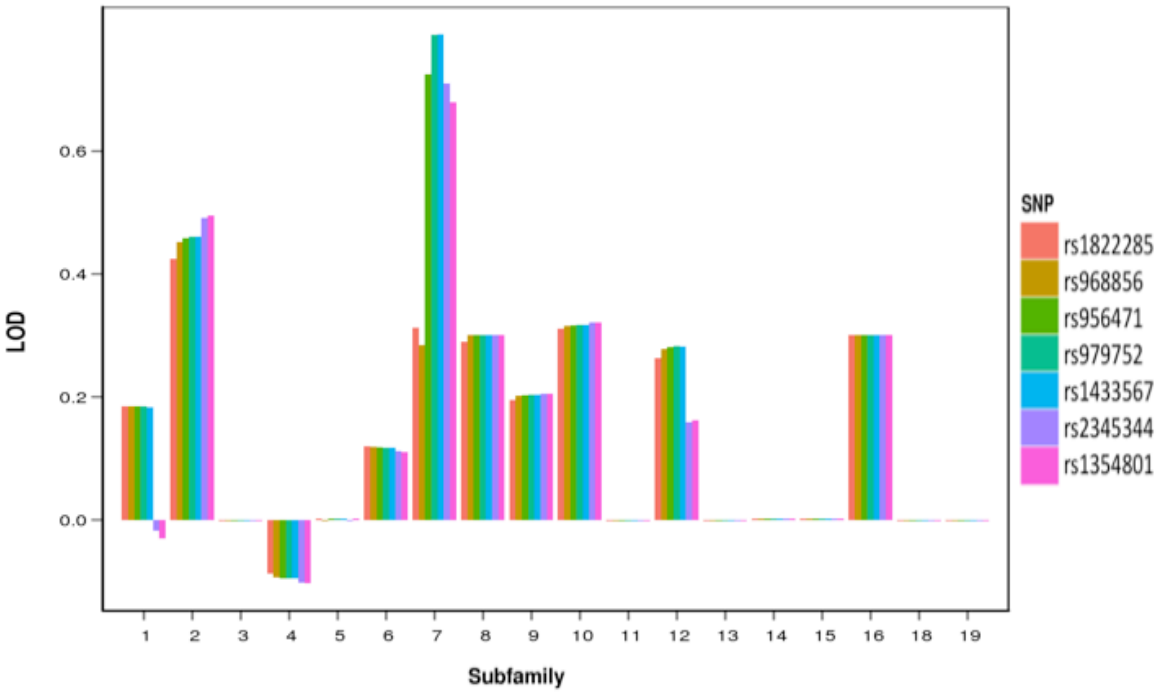


Figure 5.27. Subfamily contribution to the NPL_{all} LOD scores on chromosome 11p15.4.Subfamilies 1 through to 12 constitute Branch 1, subfamilies 13 and 14 constitute Branch 2 and subfamilies 15 through to 19 constitute Branch 3. Subfamily 17 was dropped from the analyses.

5.5.5.2 The Depression Phenotype Model: Parametric Linkage

The region on chromosome 11p15.4 is the only NPL region corroborated with some linkage evidence from parametric analyses using MERLIN. Under a dominant mode of disease transmission and using the depression phenotype model, this region achieved a maximum HLOD score of 1.72 (α=1) in Branch 1 and a maximum HLOD score of 1.78 (α=1) in the BBF. The region did not yield any LOD scores greater or equal to 1.0 under the recessive mode of transmission nor was it supported by any of the McLinkage analyses.

5.5.5.3 The Narrow, Broad, and Super Phenotype Models

The linkage region identified on chromosome 11p15.4 yielded LOD scores that did not meet criteria for suggestive or significant evidence for linkage using the narrow, broad, or super phenotype models in NPL_{all} and NPL_{pairs} tests. Therefore the results

implicate chromosome 11p15.4 as a susceptibility locus for phenotypes identified under the depression model (MDD and dysthymia) only.

5.5.6 The X-Chromosome

NPL multipoint MINX (MERLIN in X) linkage analysis was performed on subfamilies from Branch 1 (n=269) and the BBF (400) using 230 SNPs genotyped on the X chromosome using the Affymetrix 10K array. The information content, derived from the genotypes generated across the X chromosome ranged from 62 to 80%. Little evidence was observed for linkage on the X-chromosome by any phenotypic model (Figure 5.28).

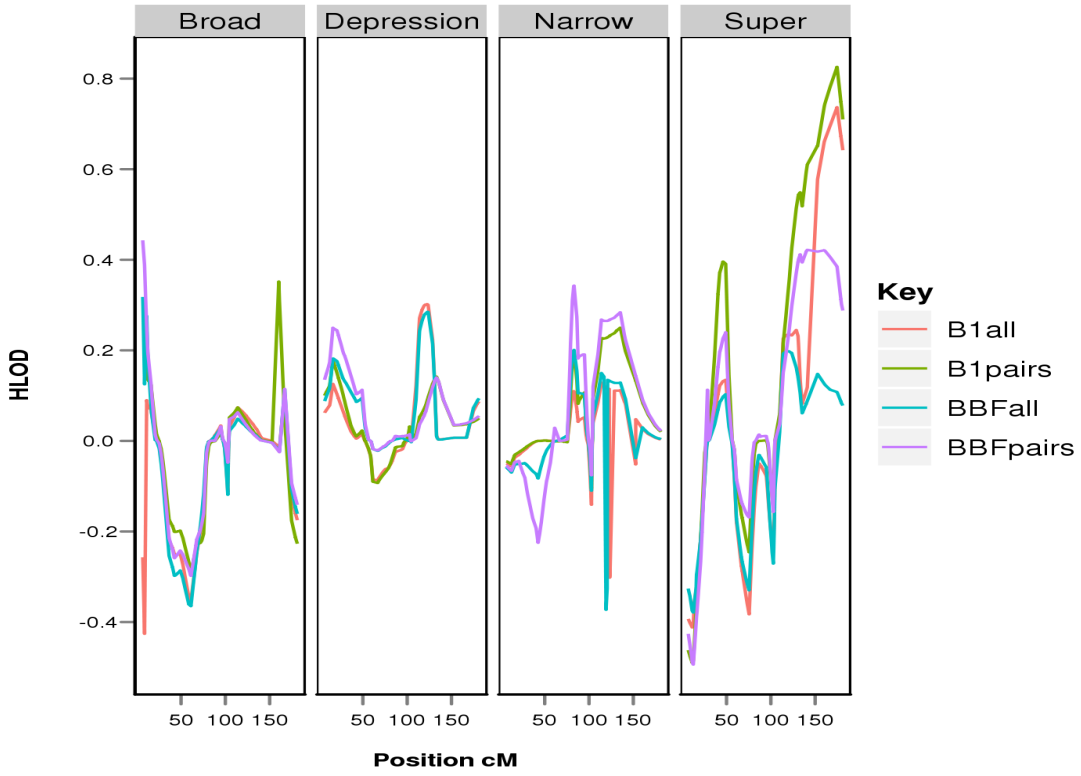


Figure 5.28. Graph showing NPL_{all} and NPL_{pairs} exponential LOD scores on the X-chromosome versus centiMorgan position for Branch 1 (B1) and the BBF under the narrow, broad, super and depression phenotype models.

5.6 Interpretation of Results using the Allele Sharing Statistic Delta

Delta is the allele sharing statistic reported by MERLIN as part of the NPL output. It measures excess IBD allele sharing between pairs and groups of affected family members in the NPL_{pairs} and NPL_{all} tests respectively *(see section 4.55)*. Delta is difficult to interpret across studies, as it is not a normalised statistic. However, it has been used as a measure of effect size to compare phenotype models within studies (Breen et al., 2010). Under the null hypothesis of no linkage, excess IBD allele sharing, as measured by Delta, approaches zero. Deviation for the null indicates excess or negative allele sharing.

A maximum allele sharing score conditional on family relatedness, genotype data, and the definition of affectedness is estimated by Delta under NPL_{pairs} and NPL_{all}. In Branch 1 and the BBF the maximum allele sharing score achievable in NPL_{pairs} by all affected pairs investigated was estimated to be approximately 10 for the narrow, broad, super, and depression phenotype models. The maximum achievable allele sharing score between groups of affected family members was estimated to be approximately 10 for the narrow, 9.34 for the broad, 2.55 for the super, and 7.51 for the depression phenotype model. A region with allele sharing scores approaching the maximum possible estimated value of Delta is considered to have a major effect on disease susceptibility.

Using hierarchal phenotype models poses a challenge to the interpretation of linkage results as increased evidence for linkage in larger groups of affecteds (due to broadening the phenotype definition) could be attributed to an increase in power as opposed to true genetic differences between the phenotype groups. The allele sharing statistic Delta offers an opportunity to further evaluate the linkage regions identified

in NPL analyses under different phenotype models. However, I emphasise that the interpretation of Delta presented in this thesis is only exploratory. Discussion of Delta for each MERLIN significant linkage region is presented below.

5.6.1 Chromosome 1p

In the disease susceptibility locus identified in Branch 1 on chromosome 1p22.2-p21.2 under the narrow and broad phenotype models both of which meet criteria for genome wide suggestive evidence for linkage, a maximum Delta score of 1.8 (LOD=2.94) was achieved under the narrow phenotype model, which was reduced to a maximum Delta score of 0.94 (LOD=2.93) under the broad phenotype model. The Delta scores under the narrow phenotype model indicated this region has a modest influence on susceptibility to narrowly defined BPD (BPI, BPII, and SAD) when compared to the maximum possible estimated value of 10. However, this influence is reduced further when considering individuals with broadly defined BPD (BI, BPII, SAD, BPNOS and cyclothymia) as affected. The maximum Delta scores for the super and depression phenotype models were estimated to be 0.27 (LOD=0.17) and 0.10 (LOD=0.01) respectively, indicating a small increase in allele sharing from what would be expected under the null hypothesis of no excess sharing (i.e. Delta of zero). Even though there is equal evidence for suggestive linkage in the narrow and broad phenotype models, IBD allele sharing that is approximately halved when the definition of affectedness is broadened to include bipolar spectrum disorders, suggests the evidence for linkage in the region is probably driven by phenotypes under the narrow model.

5.6.2 Chromosome 1q

In the disease susceptibility locus identified in the BBF on chromosome 1q21.1-q21.3 under the super phenotype model, which only achieved genome wide suggestive evidence for linkage in the NPL_{pairs} analysis, a maximum Delta score of 0.95 (LOD=2.83) was achieved, indicating this region had a modest influence on susceptibility to mood disorders defined under the super phenotype model (BPI, BPII, SAD, BPNOS, cyclothymia, MDD, and dysthymia). A similar maximum Delta score of 0.94 was estimated for the narrow phenotype model, although it did not meet criteria for genome wide suggestive evidence for linkage (LOD=0.95). The maximum Delta scores for the broad and depression phenotype models were estimated to be 0.68 (LOD=0.91) and 0.61 (maximum LOD=0.64) respectively, which is a small increase in allele sharing from the null of zero. A similar patter of Delta scores is observed in Branch 1, under which none of the phenotype models achieved suggestive significance. A maximum Delta score of 0.91 (LOD=0.81) was achieved under the narrow phenotype model, a maximum Delta score of 0.72 (LOD=0.72) under the broad phenotype model, a maximum Delta score of 0.87 (LOD=1.97) under the super phenotype model, and a maximum Delta score of 0.55 (LOD=0.45) under the depression phenotype model.

The modest differences in IBD allele sharing reported by Delta between the super phenotype model in comparison with the other phenotype models suggest that linkage evidence was not achieved using the other phenotype models possibly due to a lack of power to detect linkage rather than do to underlying differences between the phenotype groups.

5.6.3 Chromosome 2p

In the disease susceptibility locus identified on chromosome 2p23.1-p22.3 under the super phenotype model with genome-wide significant linkage using the NPL_{pairs} analysis only, maximum Delta scores of 1.10 (LOD=3.83) and 1.01 (LOD=3.46) were reported in Branch 1 and the BBF respectively, indicating this region had a modest influence on susceptibility to mood disorders defined under the super phenotype model (BPI, BPII, SAD, BPNOS, cyclothymia, MDD, and dysthymia).

The maximum Delta scores reported in Branch 1 for the narrow, broad, and depression phenotype models were 0.98 (LOD = 0.99), 0.76 (LOD=0.73), and 0.89 (LOD=1.34) respectively. The Delta scores in the BBF indicated similar allele sharing patterns across the groups in this region, with maximum scores of 0.79 (LOD=0.67), 0.58 (LOD=0.40), and 0.72 (LOD=0.47) in the narrow, broad, and depression phenotype models respectively. These Delta scores indicate the linkage evidence was not achieved using the narrow, broad, and depression phenotype models in this region possibly due to a lack of power to detect linkage rather than true differences between the phenotype groups.

Of interest, evaluation of allele sharing under the NPL_{all} analyses, under which the super phenotype group did not achieve significance, allele sharing estimated by Delta in Branch 1 was the lowest under the super phenotype model (maximum Delta=0.20, LOD= 0.80) in comparison with that reported under the narrow phenotype model (Delta=0.64, LOD=0.88), the broad phenotype model (maximum Delta = 0.43 LOD=0.42), and the depression phenotype model (maximum Delta=0.43, LOD=0.68). A similar pattern of results was reported in the BBF.

5.6.4 Chromosome 11p

In the disease susceptibility locus identified on chromosome 11p15.4 under the depression phenotype model, which met criteria for genome wide significant linkage in Branch 1 and the BBF, a maximum NPL_{all} Delta score of 1.41 (LOD=3.08) and 1.48 (LOD=3.30) were achieved in Branch 1 and the BBF respectively, indicating this region has a modest influence on susceptibility to unipolar depression. A small increase in allele sharing is reported in the narrow and super phenotype models with maximum Delta scores of 0.21 (LOD=0.04) and 0.28 (LOD=0.11) reported under the narrow phenotype model in Branch 1 and the BBF respectively and maximum Delta scores of 0.13 (LOD= 0.09) and 0.11 (LOD= 0.05) reported under the super phenotype model in Branch 1 and the BBF respectively. Further, Delta scores in the broad phenotype group were indicative of no excess sharing with zero IBD allele sharing reported in Branch 1 (LOD= -0.29) and the BBF (LOD=-0.35). A similar pattern is observed across the phenotype group in NPL_{pairs} Delta scores (data not presented here). These Delta scores suggest the region on chromosome 11p15.4 may be specific to depression only.

Chapter 6 Case Control Exploration of Linkage Regions

6.1 Introduction

Genome wide significant and suggestive linkage signals were reported in eight chromosomal regions in the BBF linkage study *(see Chapter 5 section 5.3).* To explore the generalisability of the linkage findings, two case-control association analyses of bipolar and depression case-control cohorts were performed. The aim was to clarify the role of the identified linkage regions in the general aetiology of sporadic cases of BPD and unipolar depression and in particular to explore whether there are association signals in the linkage regions that might help explain and refine the source of the linkage signals.

6.1.1 Linkage disequilibrium and Association Mapping

As described in chapter 1 *(section 1.42)* linkage disequilibrium (LD) is based on the non-random assortment between alleles in the population. Allelic association is the result of a mutation that occurred on an ancestral haplotype that was passed down from generation to generation along with the disease mutation. A number of processes contribute to the presence of LD and influence the level of LD across the genome, including rate of mutation, recombination and genetic drift (Ardlie, Kruglyak, & Seielstad, 2002). The most important of these is recombination. Over many generations, ancestral chromosomes have become mixed together, with small sections of DNA from each chromosome interspersed. As the sections of the ancestral chromosomes become smaller, the chance of a recombination event occurring to break up the section also becomes smaller. Markers in these sections are in high LD with each other and are always inherited together. LD based association mapping between cases and controls allows high resolution mapping because for many inherited diseases recombination events over the numerous generations between the mutated ancestor and current population have narrowed the DNA markers that remain part of the disease associated

haplotype to 50 kb or less (Ennis, Maniatis, & Collins, 2001), which allows for the identification of susceptibility genes with small effect sizes.

Association mapping studies are conducted to investigate polymorphisms suspected of being implicated in a disease, positional candidate genes from prior linkage studies, functional candidate genes, as well as to perform fine mapping in regions of 1 to 20 Mb identified by linkage studies, and; more recently to perform genome wide association mapping aimed at identifying common causal variants across the genome. All of these methods investigate common genetic variants, i.e. polymorphisms that occur in more than 1 percent of the population in cases versus controls, and are all sensitive to population stratification.

6.1.2 Methods of Association Mapping

Single SNP analyses in cases and controls test the null hypothesis of no association between the disease and genotypes under investigation. A number of options are available for SNP association tests. In the first, the genotype counts of AA, AB, and BB are tested in cases and controls using an observed versus expected test statistic without providing any order across the genotypes AA, AB, and BB, a so called general genetic model that typically tests the observed number of AA genotypes in cases and controls. The data may be also analysed assuming a specified genetic model. For example, for a single SNP with alleles A and B and the hypothesis that carrying allele B increased risk of disease, the data may be analysed assuming a dominant model in which carrying alleles AB and BB confer similar risk to disease or assuming a recessive model in which only carrying minor alleles BB confers risk to disease. Alternatively the data may be analysed assuming an additive genetic model in which heterozygote risk AB is intermediate between the two homozygote risks (BB > AB > AA) (Lewis, 2002). For complex diseases, risk from individuals SNPs is often considered to be additive and most association studies perform additive genetic tests unless the data

suggests the presence of dominant or recessive genetic effects. Generally, however, there is no accepted answer as to which single SNP test to perform. Adopting an additive model sacrifices power to detect association if the genetic risk is far from additive but gives greater power for near additive risk (Balding, 2006). An alternative approach is to analyse case control data by alleles, as opposed to genotypes. Allelic association tests break down the genotypes to compare the total number of A and B alleles in cases and controls regardless of genotype. This method has more power than the additive genetic model, however, it is not recommended as both cases and controls need to be in Hardy Weinberg Equilibrium (Lewis, 2002).

In addition to analysing single markers for association with a trait, a number of markers in combination can be studied. This is known as haplotype analysis. A haplotype is made up of a specific combination of alleles at adjacent SNPs (Figure 6.1). In the example shown, the alleles at three SNPs in three individuals can be seen. When the identical sequence surrounding the alleles is removed, the alleles can be combined to form a haplotype.

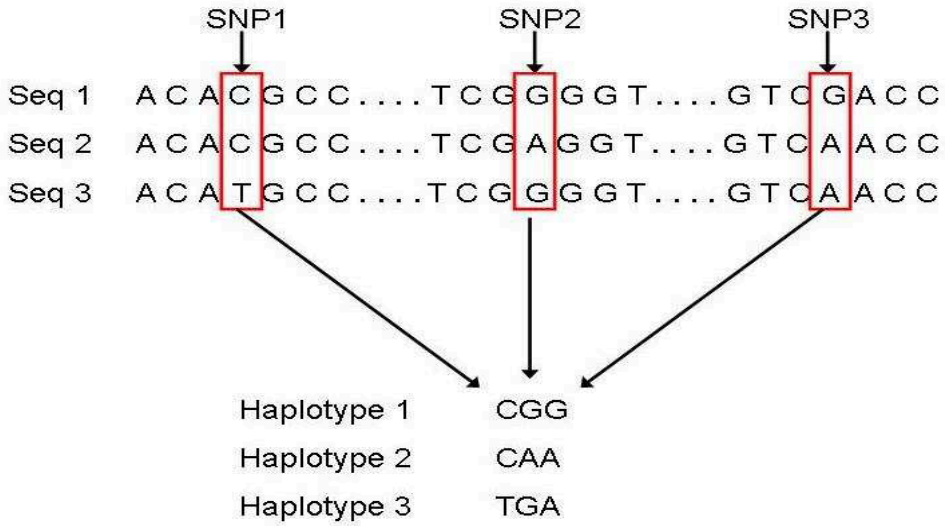


Figure 6.1. A haplotype is made up of particular combination of alleles at nearby SNPs. In the above example, there are three biallelic SNPs, highlighted by the red rectangles. The alleles of these SNPs are shown in three different individuals, surrounded by sequence , which is identical across all. The haplotype is made up by combining the alleles across each sequence, and removing the identical sequence. Adapted from www.hapmap.org.

Markers genotyped in an association study are not necessarily directly contributing to disease, but may be in LD with markers that are. For example, in Figure 6.1, assume that only SNP1 and SNP3 have been genotyped. If the A allele at SNP2 is the risk allele, this can be detected using haplotype analysis. The A allele at SNP2 is only present with a C allele at SNP1 and an A allele at SNP3. Therefore, the CA haplotype of SNP1 and SNP3 'tags' the A allele at SNP2. In this way un-genotyped susceptibility alleles for complex diseases can be mapped through their LD with genotyped marker alleles. The pattern of LD in the human genome has been extensively studied and is a great tool for haplotype analysis.

Recombination happens at different rates across the genome, regions with low recombination and high LD (haploblocks) are interspersed with regions of high recombination and low LD, so called recombination hot spots (Figure 6.2). Where blocks of LD encompassing a number of SNPs are present over a gene, the number of SNPs to be genotyped can be reduced. SNPs are chosen which predict the genotypes of a number of other SNPs in the block.

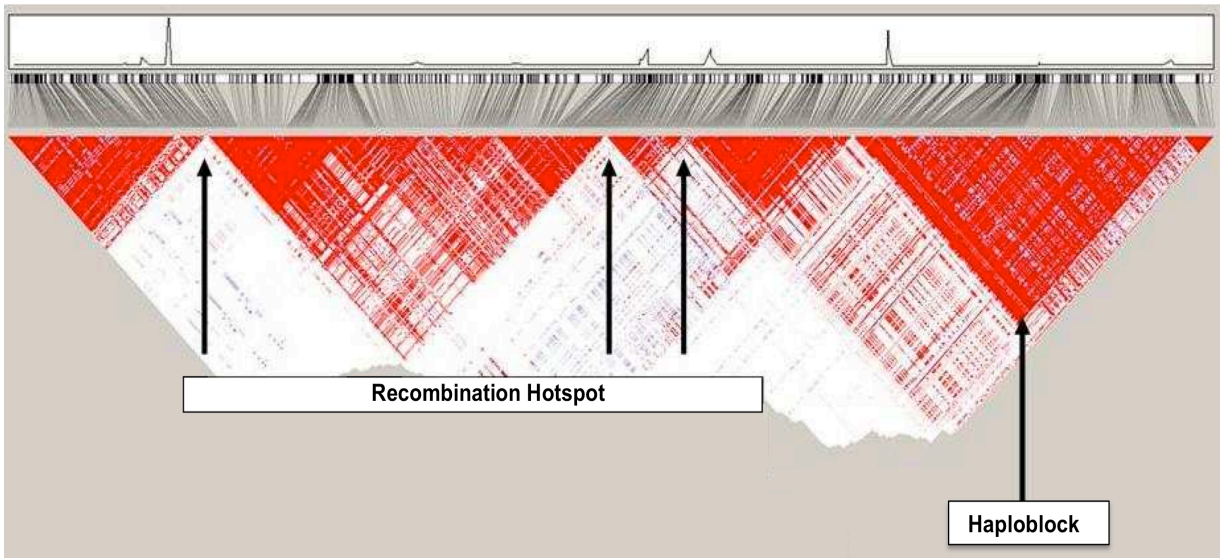


Figure 6.2 Pattern of LD across a genomic region in a specific population. Red colour represents a high level of LD between the markers, which are represented in chromosomal order at the top of the diagram. Regions of high LD are known as haploblocks. Haploblocks are split up by regions of low LD and a high degree of recombination, known as recombination hotspots. The top track represents a measure of recombination, with peaks indicating recombination hotspots.

6.1.3 Population Stratification

Association between a disease and a genetic marker can arise as an artefact of population stratification rather than genuine association to disease. Population stratification may either be caused by historically distinct subgroups (each with its own different set of allele frequencies) or intra-population heterogeneity in allele frequency due to recent admixture of historically distinct subpopulations. The problem arises when cases and controls are sampled from genetic subgroup with alleles frequencies that substantially differ irrespective of disease status resulting in spurious associations between genotype and phenotype at markers that are

completely unlinked to the disease. For example, in a population that is a mixture of African and Caucasians cases of hypertension will occur disproportionately among African Americans who have a higher prevalence of the disease. If we conduct an association study using cases of African American ancestry and Caucasians any alleles that occur more commonly in African Americans will tend to be associated with disease, even if they are completely unlinked to disease causing genes (Reich & Goldstein, 2001). In addition to confounding from false positive associations, population stratification my lead to reduced statistical power resulting in fewer true positive findings. Population stratification could be dealt with by matching cases and controls for genetic ancestry, which has become a common practice in quality control of most genome wide association studies (GWAS).

6.2 Subjects

6.2.1 Bipolar Case Control Analysis

Nine hundred and thirteen individuals with BPD, recruited from London and Toronto were used for the case-control association analysis. Details of the sample have been previously published (Gaysina et al., 2009). The mean age for Canadian subjects was 45.99 years (SD±12.54 years; n=383 total, 143 men and 240 women), and for UK subjects 48.02 years (SD±11.40;n=514 total, 176 men and 338 women). Subjects were recruited following identical protocols and were identified from psychiatric clinics, hospitals, primary care physicians, patient support groups, and from respondents to media advertisements. Subjects were included if they were over the age of 18 and had been diagnosed with BPI or BPII disorder as defined by the DSM-IV (American Psychiatric Association, 1994) or the International Classification of Disease 10^{th} edition operational criteria (ICD-10) (World Health Organization, 1993). All subjects were interviewed using the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) (Wing et al., 1990). The age of onset of disease was 21.28 ± 10.48 years. To minimize population stratification all subjects were white and of European parentage. Exclusion criteria were:

1. First degree relative having fulfilled criteria for schizophrenia.
2. Being related to an individual already included in the study.
3. Psychotic symptoms that were mood incongruent or present when there was no evidence for mood disorders.
4. Intravenous drug use with a lifetime diagnosis of drug dependence.
5. Mania or depression occurred solely as a consequence of alcohol or substance abuse/dependence and or medical illness or in response to medication.

A total of 1687 control individuals (721 men and 966 women; mean age ± SD: 42.07±13.17) were recruited from two sites; London, UK and Toronto, Canada. The mean age of the Canadian control group was 43.66 years (SD±13.12 years; n=311 total, 142 men and 169 women), and for the UK control group 41.70 (SD±13.16 years; n=1376 total, 579 men and 797 women). Control subjects from both centres were screened for lifetime absence of psychiatric disorder using a modified version of the Past History Schedule (McGuffin, Katz, & Aldrich, 1986). They were either interviewed by telephone or in person. All control subjects were white and of European parentage. Exclusion criteria were:

1. If they scored 10 or more on the Beck Depression Inventory (Beck & Steer, 1984)
2. Did not consent to give blood or check swab samples

All subjects gave written informed consent and the local ethical committees of all centres involved approved the study.

6.2.2 Depression Case-Control Analysis

A total of 2948 individuals (871 men, 2077 women; mean age ± SD: 45.25 ±12.15 years) sourced from three studies: the Depression Case-Control (DeCC) study (Gaysina et al., 2008), Depression Network (DeNT) study (Farmer et al., 2004) and the Genome-Based Therapeutic Drugs for Depression (GENDEP) study were used in the depression case-control association analysis conducted (Uher et al., 2009). The DeCC sample comprises subjects with recurrent unipolar depression (minimum two episodes) of at least moderate severity as defined by the DSM IV or ICD10 criteria, recruited from three UK sites (London, Cardiff and Birmingham). The DeNT is an affected sibling pair linkage study consisting of cases with recurrent unipolar depression of at least moderate severity recruited form nine European sites and one US site. Probands only were used for the association analyses. The GENDEP study includes individuals with one episode of depression of at least moderate severity recruited

from nine European centres. The depression in the later study was not necessarily recurrent. As with the Bipolar Case Control subjects, all subjects were interviewed using the SCAN. To minimize population stratification all subjects were white and of European parentage. Exclusion criteria were:

1. First degree relative having fulfilled criteria for mania, hypomanic, schizophrenia, or BP disorder.
2. Depression occurred solely as a consequence of alcohol or substance abuse/dependence, a medical illness, or in response to medication.

A total of 1288 control individuals contacted via the Medical Research Council general practice research framework were interviewed by telephone using the Past History Schedule (McGuffin et al., 1986). Of these subjects, 58.4% were women, with mean age of 47.24 years (range=20–69). A further 457 healthy volunteers (61.4% female) were staff or students of King's College London, again screened for mental health using the Past History Schedule. To minimize population stratification all subjects were white and of European parentage. Details of the sample and the methods used to assess the subject have been previously published (Farmer et al., 2004; Lewis et al., 2010).

6.3 Materials and Methods

6.3.1 Case-Control Genotype Data

Cases and controls from the bipolar and depression case control studies were genotyped using the Illumina HumanHap610-Quad BeadChips by the Centre National de Génotypage. The data underwent stringent quality control procedures by members of the SGDP research team for previous studies. Briefly, subjects were excluded if their genotypic data showed a missing rate greater than 1%, abnormal heterozygosity, or a sex assignment that conflicted with phenotypic data. SNPs with a missing rate greater than 1%, minor allele frequency less than 1%, or showing departure from Hardy Weinberg equilibrium ($p < 1\times10^{-5}$) were also excluded (Lewis et al., 2010). In addition, subjects who failed PLINK pair-wise Identity-by-Descent (IBD) estimation test (--genome) (*see Chapter 4 section 4.4.2)* by having an IBD pair-wise estimate greater than 0.05, which indicates relatedness to a subject already included in the study were removed. After quality control there was a total of 502,877 SNPs in the bipolar case control study and 471,581 SNPs in the depression case control study.

6.3.2 Statistical Methods

6.3.2.1 Extraction of Linkage Regions Identified Under the BPD Phenotype

SNPs identified with significant or suggestive evidence for linkage in the BBF linkage analysis using BBF members diagnosed with BPI, BPII, SAD, BPNOS, and cyclothymia as affected (i.e. under the narrow, broad, and super phenotype models) were extracted from the HumanHap610-Quad BeadChips SNP data using the PLINK --extract command for inclusion in the bipolar case-control analysis. In total 1723 SNPs were available for analysis on chromosome 1p22.2-p21.2, 1046 SNPs on chromosome 1q21.1-q21.3, 1133 SNPs on

chromosome 2p23.1-p22.3, 1814 SNPs on chromosome 3p24.3-p24.1, and 2175 SNPs on chromosome 22q11.21-q12.1.

6.3.2.2 Extraction of Linkage regions Identified Under the Depression Phenotype

The linkage regions identified with significant or suggestive evidence for linkage in the BBF linkage analyses using BBF members diagnosed with depression as affected (i.e. depression phenotype and super phenotype models) were included in the depression case-control analysis. In total 930 SNPs were available for analysis on chromosome 1q21.1-q21.3, 1002 SNPs on chromosome 2p23.1-p22.3, 756 SNPs on chromosome 11p15.4, 2002 SNPs on chromosome 12p13.32-p13.31, and 1901 SNPs on chromosome 12q24.22-q24.32.

6.3.2.3 Ancestry Principal Component Analysis

As discussed, population stratification is problematic in case control association studies as ancestry differences between cases and controls can lead to false positive, or false negative association between phenotype and genotype (Marchini, Cardon, Phillips, & Donnelly, 2004). To account for population stratification in the bipolar and depression case control cohorts, principal component analysis (PCA) was conducted.

PCA is a multivariate technique in which a number of related variables are transformed into sets of uncorrelated variables (Jackson & Wiley, 1991). Multivariate analyses involve observation and analyses of more than one statistical variable at a time. Fundamentally, PCA aims to reduce the number of variables, revealing the underlying structure of the dataset explaining the majority of the variance of the data (Dunteman, 1989; Manly, 2005). PCA is non-parametric meaning no assumption about the data is made (Jolliffe & MyiLibrary, 2002). It is often used for exploratory data analysis where the underlying structure of a dataset is unknown. PCA seeks a linear combination of variables, such that the maximum variance is extracted. It then removes this variance and seeks a second linear

combination, which explains the maximum proportion of the remaining variance, and so on. The maximum number of components that can be extracted from a PCA is equal to the number of variables in the data. However the aim is to extract a smaller number of components that explain nearly all of the variance present in all of the original data (Jolliffe & MyiLibrary, 2002). The first component explains the majority of variance observed in a dataset and each subsequent component explains a declining amount of the remaining variance. Usually, only the first few components account for a meaningful amounts of variance.

The scree plot is one way to determine the number of components that explain the majority of the variance. It plots the principal components as the X-axis and the corresponding eigenvalues (the amount of variance explained) as the Y-axis (Cattell, 1966). As one moves to the right, toward later components the eigenvalues drop. The place where the smooth decrease of eigenvalues levels off to a less steep decline (making an elbow) denotes the position where further components no longer contribute to the overall variance and should be dropped. This technique is problematic when there is no clear break or multiple break points exist in the scree plot.

The program EIGENSOFT (Price et al., 2006) was used to derive ancestry principal components for incorporation into the association analyses. The program uses PCA to model ancestry differences between cases and controls using SNP information. To avoid confounding the PCA, the bipolar and depression datasets were pruned using PLINK linkage disequilibrium based SNP pruning. A window size of 1500 and a genotypic correlation (r^2) of 0.2 were specified. Using these parameters, the program calculated LD between each pair of SNPs within the specified window and removed one of the SNP pairs if the LD or genotypic correlation (r^2) between them was greater than 0.2. EIGENSOFT PCA was conducted on the pruned datasets.

6.3.2.4 Correction for Multiple Testing

Permutation tests are statistical randomisation tests that rearrange labels on the observed data many times under the assumption of no association, calculating the test statistic that each re-arrangement yields. Labels of cases and controls are interchanged under the null hypothesis of no association so only the phenotype-genotype relationship is changed, while all aspects of the data (e.g. genotypes, LD structure) are maintained. This is repeated multiple times determined by the user and the empirical p-value is generated using the formula:

$$P_{perm} = \frac{(r+1)}{(n+1)}$$

Where P is the corrected p-value following the permutation procedure, r is the number of times the permuted tests exceeded the observed statistic obtained from the real dataset, and n is the total number of permutations (North, Curtis, & Sham, 2002). Permutation testing is a computationally demanding task. The p-value attained from the actual test statistic is what should determine the number of permutations required. Should a low p-value be attained by the actual data (e.g. $p=10^{-6}$) then the number of permutation runs should reflect this (10^6 or 10,000,000, permutations). However, generally, in genetic studies 10,000 permutations is considered acceptable as p-values are commonly more modest (equal to or less than 10^{-4}) (Gao, Becker, Becker, Starmer, & Province, 2010).

6.3.2.5 Bipolar Case Control Association Tests

PLINK logistic regression tests for association with 913 bipolar cases and 1687 screened comparison controls at the four chromosomal regions identified by the BBF linkage study under the narrow, broad, and super phenotype models using ancestry principal components as covariates were conducted, assuming an additive genetic model for all the regions under

study. In addition, association tests under a dominant model of disease transmission in region 3p25-24 and a recessive model of disease transmission in region 22q11.21-q12.1 were performed, as informed by findings from the BBF linkage study. Correction for multiple testing was applied using 10,000 multiple permutation (--mperm) of the dataset.

6.3.2.6 Logistic Regression Depression Case Controsl Analysis

PLINK logistic regression tests for association with 2948 depression cases and 1288 screened comparison controls at the three chromosomal regions identified by the BBF linkage study under the depression and super phenotype models using ancestry principal components as covariates. The association analyses were conducted assuming an additive genetic model for all chromosomal regions under study, as well as a dominant genetic model in regions 11p15.4, 12p13.32-p13.31, and 12q24.22-q24.32 as informed by findings from the BBF linkage study. Correction for multiple testing was applied using 10,000 multiple permutation (--mperm) of the dataset.

6.4 Results

6.4.1 Bipolar Case Control Analyses

6.4.1.1 Principal Component Analysis in the Bipolar Case Control Sample

Pruning the bipolar case control dataset resulted in 101513 SNPs that were in approximate linkage equilibrium with each other. The pruned data set was used in the analysis. Twenty-three individuals were outliers and were therefore removed from the dataset. Based on the scree plot (Figure 6.3), the first four principal components were used as covariates in the association analysis. The clustering of the bipolar cases and controls by ancestry, as indicated by the first four principal components, is shown in Figure 6.4.

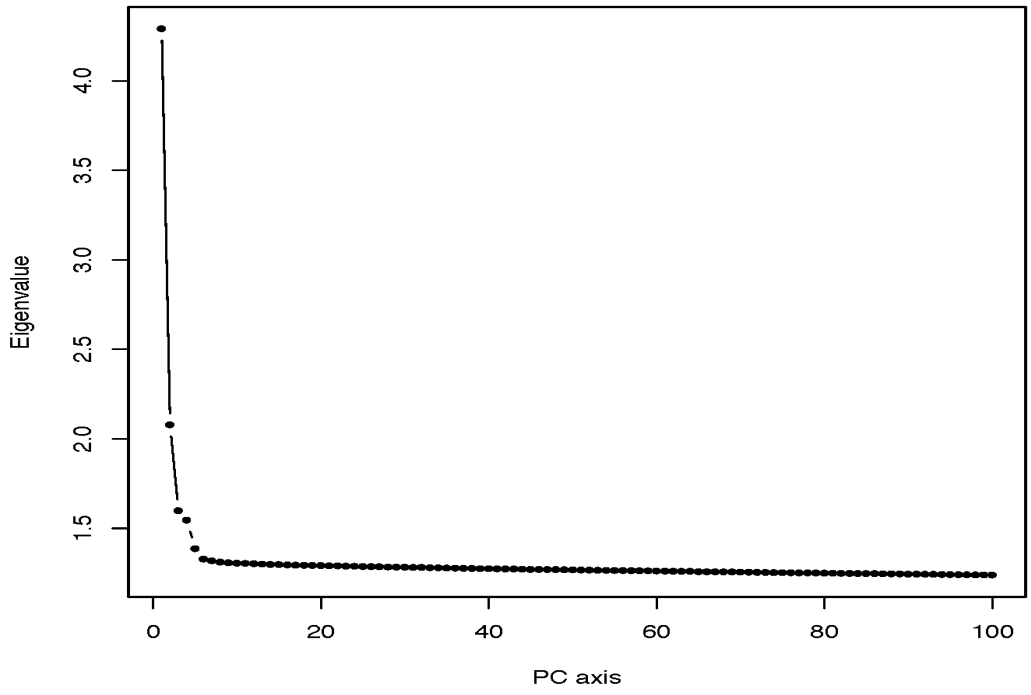


Figure 6.3. Scree plot for the bipolar case-control study showing the principal components (PC) across the x-axis and their respective eigenvalues on the y-axis. The scree plot shows that four components explain the majority of the variance in the data.

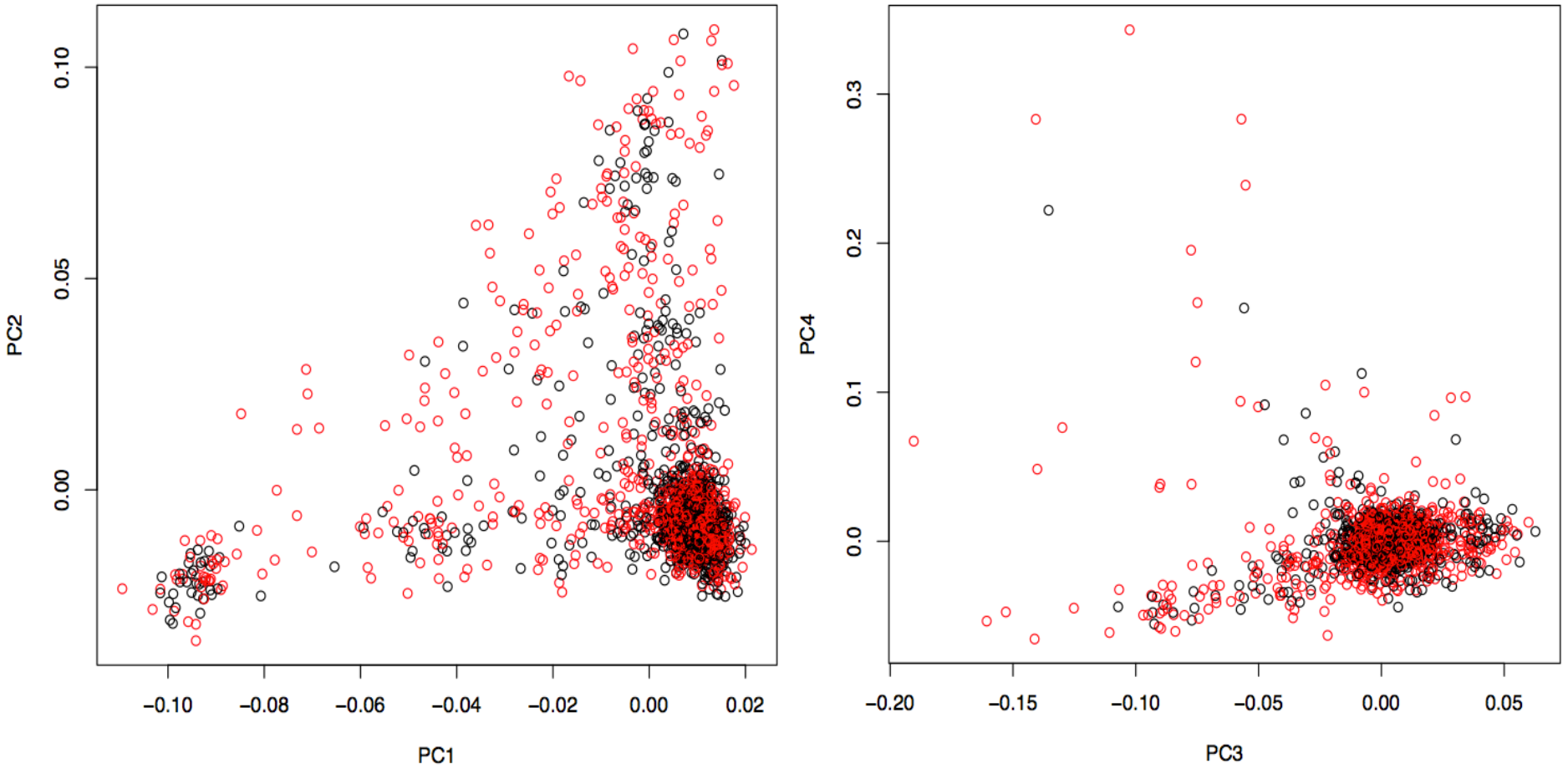


Figure 6.4. The first four principal components (PC) for the bipolar case control data after the removal of outliers (23 individuals). The graph shows the pattern of clustering according to ancestry in the cases (black) and controls (red). These first four factors, which explain the majority of the variance in ancestry were included in the case-control association analysis to account for population stratification.

6.4.1.2 Logistic Regression Bipolar Case Control Analysis

The results of the association analysis showed none of the SNPs in any of the interrogated regions achieved significance under the additive or other models indicated by significant or suggestive linkage signals in the BBF analysis. The top two SNPs from each region are presented in Table 6.1.

Chromosomal Region	Model	SNP	Physical Position	Odds Ratio	Minor Allele	Corrected P-value
1p22.2-p21.2	ADD	rs2497750	99119372	1.295	T	0.4265
		rs524322	94253345	1.232	C	0.9389
1q21.1-q21.3	ADD	rs12047833	150190059	0.6908	C	0.1922
		rs3006423	150214737	0.6647	C	0.1953
2p23.1-p22.3	ADD	rs1366817	31504054	1.214	G	0.9413
		rs13404913	33520608	0.7982	C	0.9511
3p25.3-p24.1	ADD	rs3860582	21616961	0.7618	T	0.1045
		rs11708571	21566381	0.7505	G	0.1399
	DOM	rs1879167	21764574	0.6673	C	0.4965
		rs3860582	21616961	0.7317	T	0.7283
22q11.21-q12.1	ADD	rs5761045	24078813	0.7081	A	0.31
		rs5996921	24078331	0.7145	G	0.3584
	REC	rs131677	21957219	1.899	C	0.5894
		rs1297593	26494075	2.335	C	0.3308

Table 6.1. Results from the bipolar case-control association analysis in regions indentified in the BBF linkage study under the narrow (1p22.2-p21.2, 3p25-24), broad (22q11.21-q12.1) and super (1q21.1-q21.3, 2p23.1-p22.3) phenotype models. Association tests were performed under the additive (ADD) genetic model in all of the regions, in addition to dominant (DOM) in the chromosome 3p25.3-p24.1 region and recessive (REC) in the chromosome 22q11.21-q12.1 region. None of the SNPs showed association with BPD.

6.4.2 Depression Case Control Analyses

6.4.2.1 Principal Component Analysis in the Depression Case Control Sample

Pruning the depression case control dataset resulted in 99240 SNPs that were in approximate linkage equilibrium with each other. The pruned data set was used in the EIGENSOFT analysis. Twenty-eight individuals were outliers and were therefore removed from the dataset. Based on the scree plot (Figure 6.5) the first four principal components were used as covariates in the association analysis. The clustering of the depression cases and controls by ancestry, as indicated by the first four principal components, is shown in Figure 6.6.

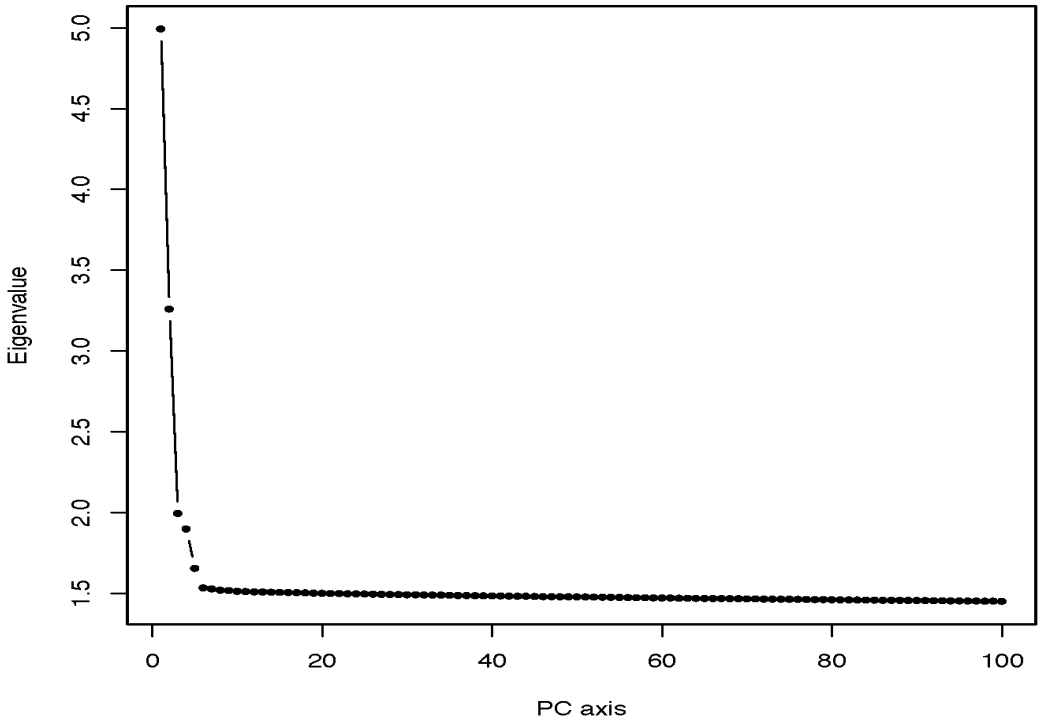


Figure 6.5. Scree plot for the depression case-control study showing the principal components (PC) across the x-axis and their respective eigenvalues on the y-axis. The scree plot shows that the first four components explain the majority of the variance in the data.

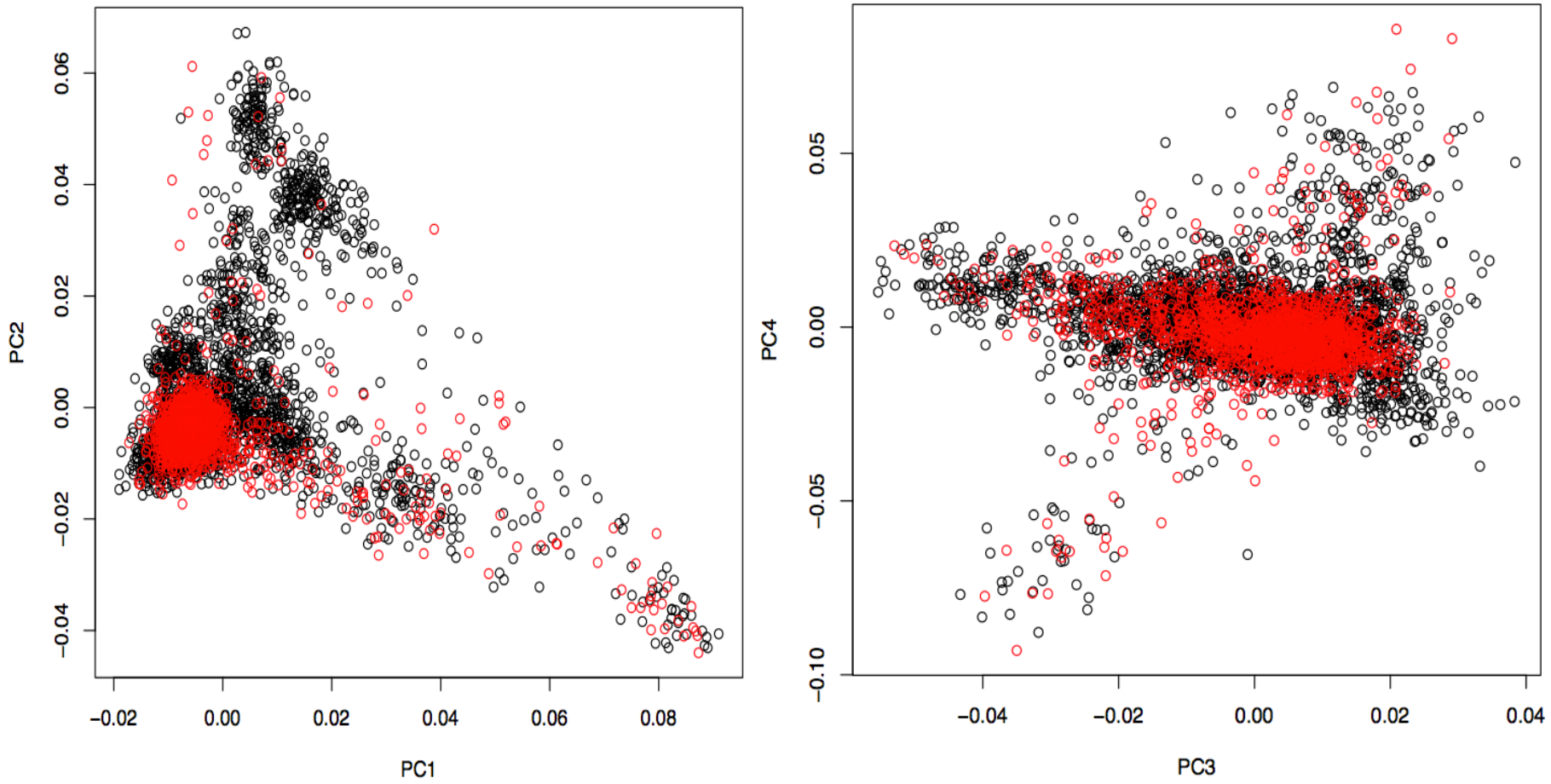


Figure 6.6. The first four principal components (PC) for the depression case control data after the removal of outliers (28 individuals). The graph shows the pattern of clustering according to ancestry in cases (black) and controls (red). These first four factors, which explain the majority of the variance in ancestry were included in the case-control association analysis to account for population stratification.

6.4.2.2 Logistic Regression Depression Case Control Analysis

The results of the association analysis showed none of the SNPs in any of the interrogated regions achieved significance under the additive or dominant models indicated by significant or suggestive linkage signals in the BBF analysis. The top two SNPs from each region are presented in Table 6.2.

Chromosomal Region	Model	SNP	Physical Position	Odds Ratio	Minor Allele	Corrected P-value
1q21.1-q21.3	ADD	rs11205347	147023193	1.149	T	0.145
		rs956796	146999164	1.152	T	0.221
2p23.1-p22.3	ADD	rs946577	153899204	1.042	G	0.55
		rs947661	153307989	0.9874	T	0.86
11p15.4	ADD	rs7934676	4355857	1.162	G	0.52
		rs4910623	4346215	0.8852	A	0.91
	DOM	rs12293167	4344336	0.7832	T	0.34
		rs4910623	4346215	0.7843	A	0.37
12p13-12	ADD	rs887357	3344906	1.258	C	0.23
		rs11046589	8699273	1.193	C	0.23
	DOM	rs12818044	493187	1.303	A	0.37
		rs2238023	2072798	0.6123	C	0.37
12q24.22-q24.32	ADD	rs830123	120910600	1.251	A	0.54
		rs7964822	118839946	0.8259	C	0.83
	DOM	rs7964822	118839946	0.7844	C	0.41
		rs830123	120910600	1.289	A	0.43

Table 6.2. Results from the depression case-control association analysis in regions identified in BBF linkage study under the depression (11p15.4, 12p13-12, and 12q24.22-q24.32) and super (1q21.1-q21.3, 2p23.1-p22.3) phenotype models. None of the SNPs showed association with depression. Association tests were performed under the additive (ADD) genetic model in all of the regions, in addition to dominant (DOM) genetic model in the chromosome 11p15.4, 12p13-12, and 12q24.22-q24.32 regions. None of the SNPs showed association with depression.

6.5 Conclusion

None of the SNPs identified in the BBF linkage regions showed an association with disease in the case-control cohorts. The results suggest that the linkage regions identified in the BBF do not play a general role in the aetiology of common forms of BPD or unipolar depression in the population. It is plausible that the linkage regions identified in the BBF are rare and family specific. However, a limitation of this study is the sample size of the case-control cohort, which may have reduced power to detect risk variants with low odds ratios. The findings of the case-control cohort will be discussed in reference to the BBF linkage findings in more detail in the general discussion of this thesis (Chapter 7).

Chapter 7 Discussion

The overarching aim of this thesis was to localise and identify genetic variants that underlie the aetiology of BPD. However, the picture emerging from the BBF linkage study is one of a general liability to mood disorders, both with linkage hits specific to either BPD or depression and hits that go across the mood disorder spectrum. When we initially collected this family it was hypothesised that this was primarily a 'bipolar' family with depression cases reflecting bipolar cases early in the development of their disorder or "milder" bipolar cases. However, the linkage results make it clear that while there are genetic risk loci segregating in the family, these act across the mood spectrum and are not preferentially transmitted to clear bipolar cases only. Given this, the family is best interpreted in genetic terms as a family with mood spectrum disorders, in which a set of loci of larger (if not large) effect segregate. This may be alongside a higher polygenic loading for these disorders.

This chapter will commence with a discussion of the results of the whole genome linkage analyses presented in chapter 5. The strengths and the limitations of the linkage method will be discussed in light of the study findings. Secondly, the results form the case-control replication study presented in chapter 6 will be discussed in terms of the generalisability of linkage findings, and the common disease/common variant and common disease/rare variant hypotheses. Finally the future directions of the BBF study will be presented

7.1 Whole Genome Linkage Regions

I successfully performed a whole genome linkage analysis on a large multigenerational family segregating severe BPD and mood disorders. After correcting for multiple testing, and from across the phenotypes tested, four regions on chromosomes 2p23.1-p22.3, 3p24.3-p24.1, 11p15.4, and 12q24.22-q24.32 achieved genome wide significance while four regions on chromosomes 1p22.2-p21.2, 1q21.1-q21.3, 12p13.32-p13.31, and 22q11.21-q12.1 achieved genome wide suggestive linkage (Figure 7.1). Many of these regions overlap with and provide further support for previous linkage findings. It is important to note that concordant linkage findings between BPD studies, as well as between BPD and other psychiatric disorder studies, do not necessarily mean that the underlying genetic susceptibility found is the same. Different genes under the same linkage peaks could be driving the linkage signals. Concordant linkage regions might also occur by chance, given the large number of chromosomal regions that are potentially involved in BPD aetiology and other psychiatric disorders (Kendler, 2006). Drawing comparisons between linkage studies should therefore be regarded as descriptive until fine mapping studies are conducted to identify the disease causing genes across studies.

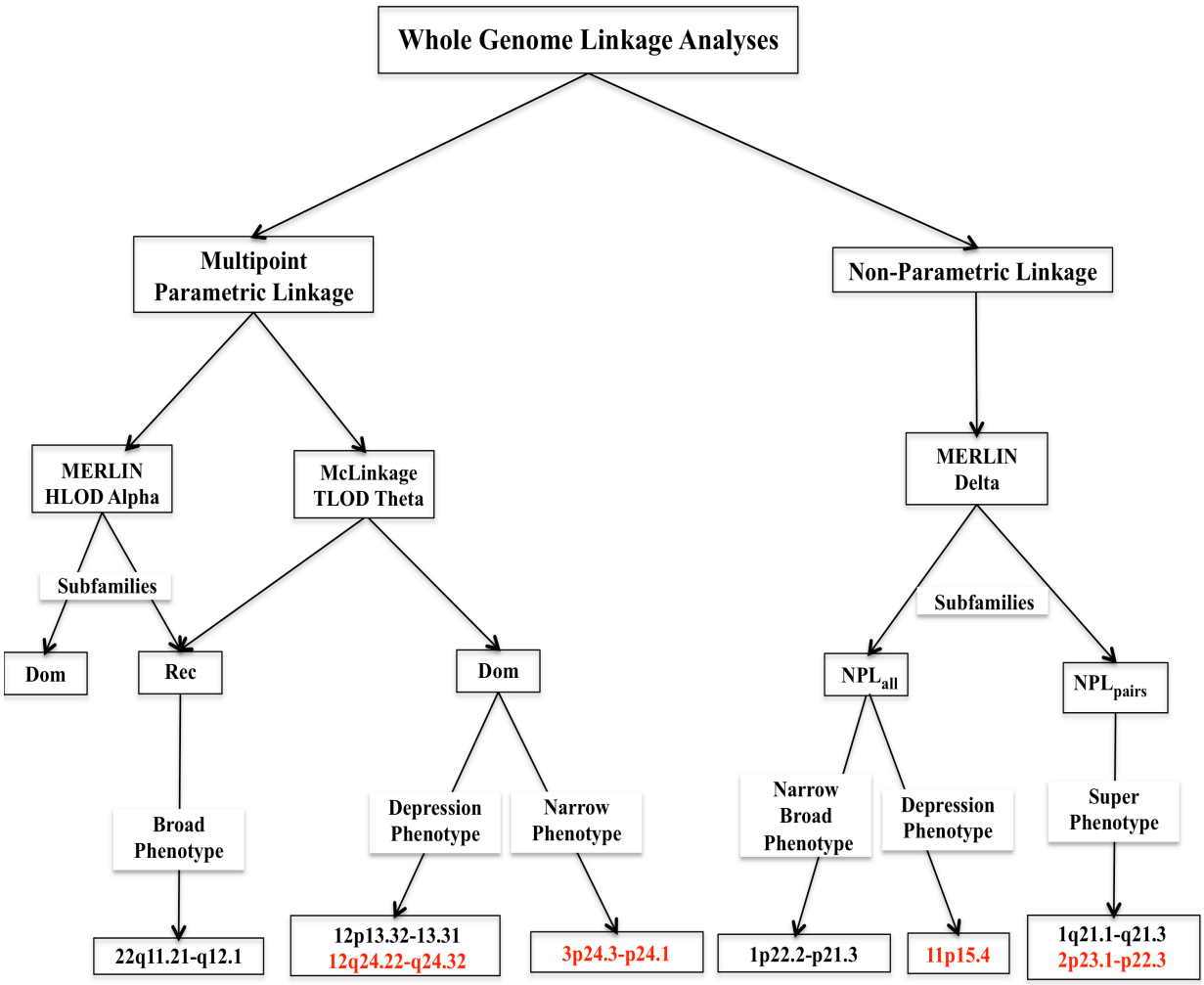


Figure 7.1. Flow chart detailing the results of the parametric and non-parametric linkage analyses performed using MERLIN and McLinkage. The four regions highlighted in red achieved genome wide significance and the four regions in black achieved genome wide suggestive linkage. Chromosomal regions were identified under all four phenotype models. Only one region on chromosome 22q11.21-q12.1 was identified using both MERLIN and McLinkage. None of the same regions reached suggestive or significant thresholds using both parametric and non-parametric linkage. In parametric linkage three regions were identified under a dominant (Dom) mode of disease transmission (3p24.3-p24.1, 12p13.32-p13.31, and 12q24.22-q24.32) and one region under a recessive (Rec) mode of disease transmission (22q11.21-q12.1).

7.1.1 Chromosome 1p

A suggestive linkage region that confers susceptibility to BPD was identified in the BBF on chromosome 1p22.2-p21.2 under the narrow (maximum LOD=2.96) and broad phenotype models (maximum LOD=2.93). This region has been implicated in previous linkage studies of BPD. It was first reported as a susceptibility locus (maximum LOD=3.4) in sibling pairs affected with BPI and BPII disorder in families from Sardinia (Del Zompo et al., 2010). An earlier genome scan in two multigenerational families from Denmark reported 1p22-21 as a suggestive linkage region for BPI, BPII, and SAD using marker D1S216 (Ewald, Flint, Kruse, & Mors, 2002). However, this marker is positioned at the 1p31.1 region according to the new build of the human genome (hg19, UCSC browser) and is approximately 11.8 Mb telomeric to the linkage peak identified in the BBF. Further, this region overlaps with a larger suggestive linkage region identified on 1p31.1-p21.2 (maximum LOD=1.32) using fifteen pedigrees from Antioquia, a population isolate in Colombia (Kremeyer et al., 2010). Their finding was, however, consistent with a general mood disorder liability including BPD and unipolar depression, as opposed to the current study that indicated region 1p22.2-p21.2 as a susceptibility locus for BPD only.

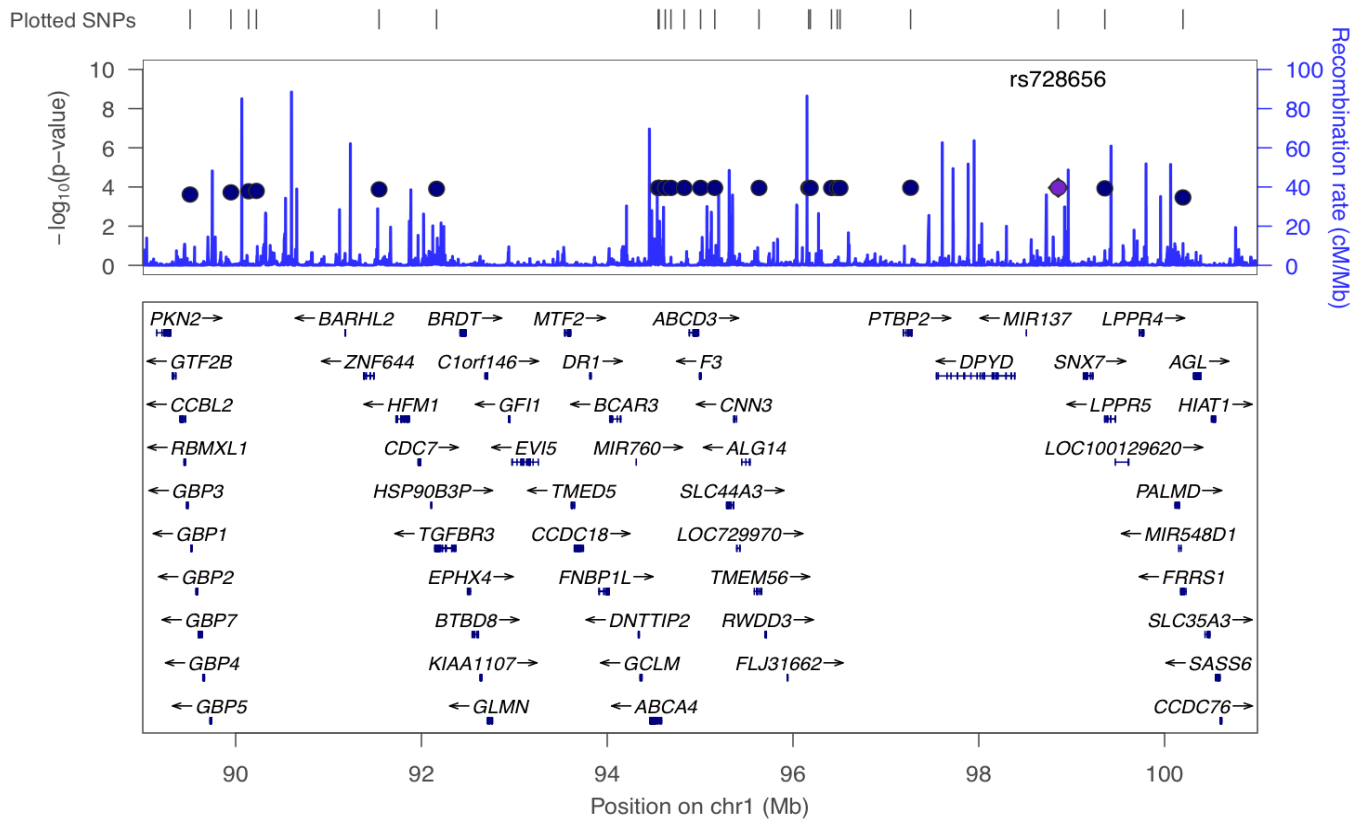


Figure 7.2. Graphical display of genes located on chromosome 1p22.2-p21.2 according to the human genome build 19. The graph depicts chromosomal positions in mega bases (Mb). The p-value for each SNP is plotted with the recombination rate between SNPs. P-values from the NPL_{all} analysis of Branch 1 under the narrow phenotype are presented in the graph. The purple diamond depicts the most significant SNP. Plotted with LocusZoom (Pruim et al., 2010)

Figure 7.2 shows, amongst the many genes in the 1-LOD region is the *GCLM* gene, which codes glutamate cysteine ligase modifier, one of three enzymes responsible for Glutathione synthesis. Glutathione is involved in the detoxification of reactive oxygen and other radical species. Its role has been investigated in BPD. Studies have shown that the mood stabilising drugs lithium and valproate increase expression of the glutathione s-transferase *(GST)* gene, that encodes enzymes known to detoxify endogenous and exogenous agents and participate in the activation and inactivation of oxidative metabolites. Fullerton et al., (2010) investigated several genes in the oxidative stress pathway, including *GCLM*, in 325 BPD patients and 392 controls and

found nominal association between BPD and genes in this metabolic pathway. The region also harbours *MIR137* (MicroRNA 137). A recent genome wide association study found a strong association ($P = 1.6 \times 10^{-11}$) between schizophrenia and rs1625579 located within an intron of a putative primary transcript for *MIR137* (Ripke et al., 2011). The role of *MIR137* in regulating neuronal maturation and adult neurogenesis has been recently reported. Its over-expression has been shown to inhibit dendritic morphogenesis, phenotypic maturation, and spine development both in brain and cultured primary neurons (Smrt et al., 2010). These finding suggests that a pathological pathway involving the dysregulation or dysfunction of *MIR137* may underlie the aetiology of schizophrenia and possibly BPD, particularly as a recent study by (Kwon, Wang, & Tsai, 2011) showed that *CACNA1C*, a confirmed BPD gene (Sklar et al., 2011) is a *MIR137* target. This further supports the involvement of *MIR137* in the aetiology of both BPD and schizophrenia. There are many additional genes with roles in the brain in this region. Nevertheless, the identification of a suggestive and replicated linkage peak harbouring *MIR137* in the BBF suggests the microRNA is possibly involved in BPD susceptibility. However, fine mapping is required to confirm its role.

7.1.2 Chromosome 1q

A suggestive linkage region was identified in the BBF genome scan on chromosome 1q21.1-q21.3 (maximum LOD=2.83). No BPD study has reported linkage in this region although one study found suggestive evidence for linkage at region 1q23.3 (approximately 6 Mb centromeric to the putative BBF region) for narrowly defined BPI and BPII disorder in 41 Ashkenazi Jewish families (Fallin et al., 2004) with a maximum LOD score of 2.46. This locus overlaps with a highly significant (HLOD=6.50) schizophrenia locus identified in 22 Celtic Canadian families at 1q21-q23 (Brzustowicz, Hodgkinson, Chow, Honer, & Bassett, 2000) with the BBF linkage peak occurring approximately 7.7 Mb more centromeric to their linkage peak.

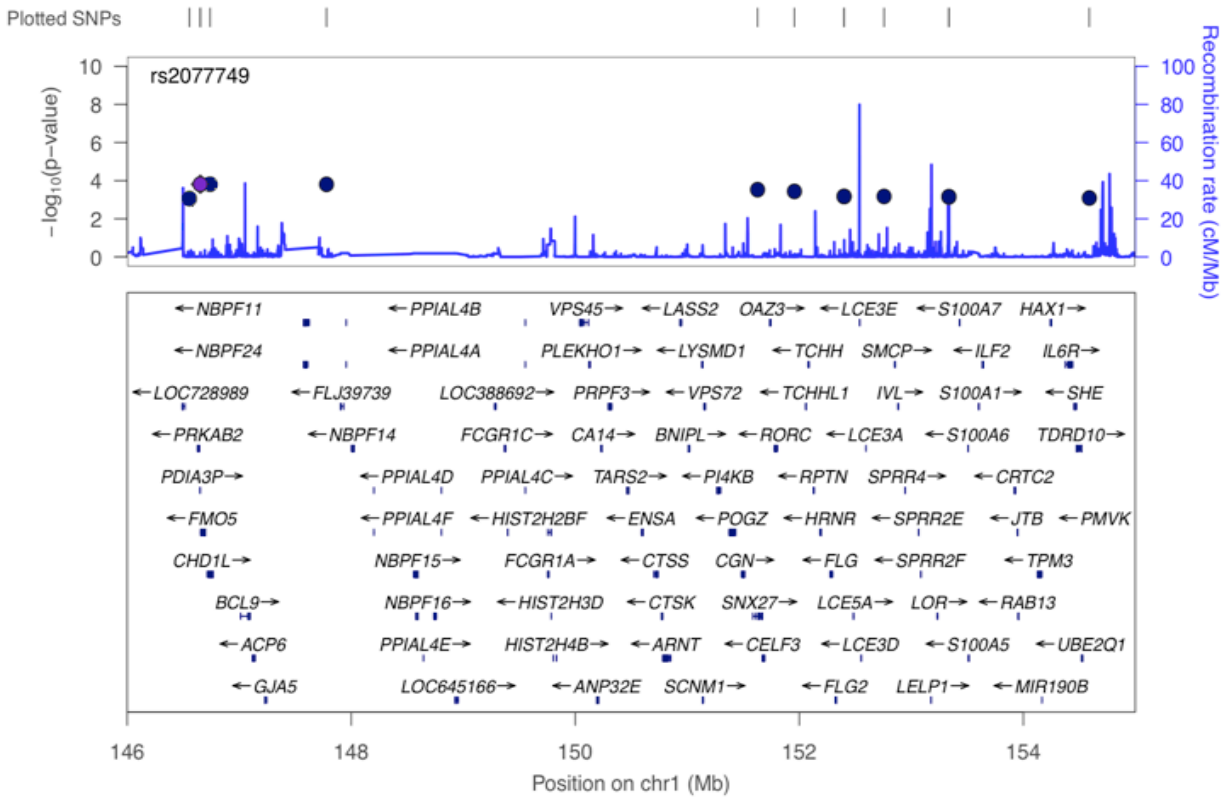


Figure 7.3. Graphical display of genes located on chromosome 1q21.1-q21.3 according to the human genome build 19. The graph depicts chromosomal positions in mega bases (Mb). The p-values for each SNP is plotted with the recombination rate between SNPs. P-values from the NPL_{pairs} analysis of the BBF under the super phenotype model are presented in the graph. The purple diamond depicts the most significant SNP. Plotted with LocusZoom (Pruim et al., 2010).

As indicated by Figure 7.3, this region has a number of genes of interest including *PRKAB2,* which is a candidate gene for diabetes mellitus (Jablonski et al., 2010; Prochazka, Farook, Ossowski, Wolford, & Bogardus, 2002). The co-morbidity between BPD and diabetes (Mcintyre, Konarski, Misener, & Kennedy, 2005), along with the prevalence of type I diabetes in the BBF, could potentially explain the suggestive linkage observed in this region. It is possible that the linkage signal is in fact due to diabetes as opposed to BPD. However, chromosome 1q21.1-q21.3 remains an interesting locus for BPD in its own right, as a family of genes encoding calcium-binding proteins, *S100A1, S100A5, S100A6,* and *S100A7* is located in this region. The S100 protein family is essential in the regulation of serotonin signalling in the brain via interactions with serotonin receptors and are implicated in depression (Svenningsson et al., 2006). Expression of the S100 gene family has also been shown to be upregulated in the frontal cortex of rats chronically treated with olanzapine, a drug typically used to treat schizophrenia, BPD, and some forms of depression, suggesting their involvement in the pathobiology of psychiatric disorders (Fatemi et al., 2006).

7.1.3 Chromosome 2p

The region on chromosome 2p23.1-p22.3 presents evidence for genome wide significant linkage in the BBF (maximum LOD=3.83), under the super phenotype model that includes both BPD and unipolar depression. It partially overlaps with a region of possible linkage implicated with nominally significant evidence by (Pato et al., 2004) in 102 Portuguese Island families with multiple patients suffering from BPD and SAD (LOD=1.13[1]) and is approximately 401 kb from another reported suggestive linkage peak on 2p23.1 (LOD=1.98) reported in 75 BPD families of German, Israeli, and Italian origin with BPI disorder only (Cichon et al., 2001). These are the only two linkage scans that support this region as a susceptibility region for BPD, however they do implicate it as a specific susceptibility locus for BPI disorder and SAD, whereas in the BBF it confers liability to mood disorders in general.

The limited support for this region in the literature could be because the finding regarding its role in general mood liability is a false positive, or alternatively it could be because the region has no significant impact on other families, but is of specific relevance to mood disorder susceptibility in the BBF.

[1] This region was reported by Pato et al. (2004) as an NPL score of 2.28 that was converted to a LOD score using LOD=NPL^2/4.62 (Ott, 1999).

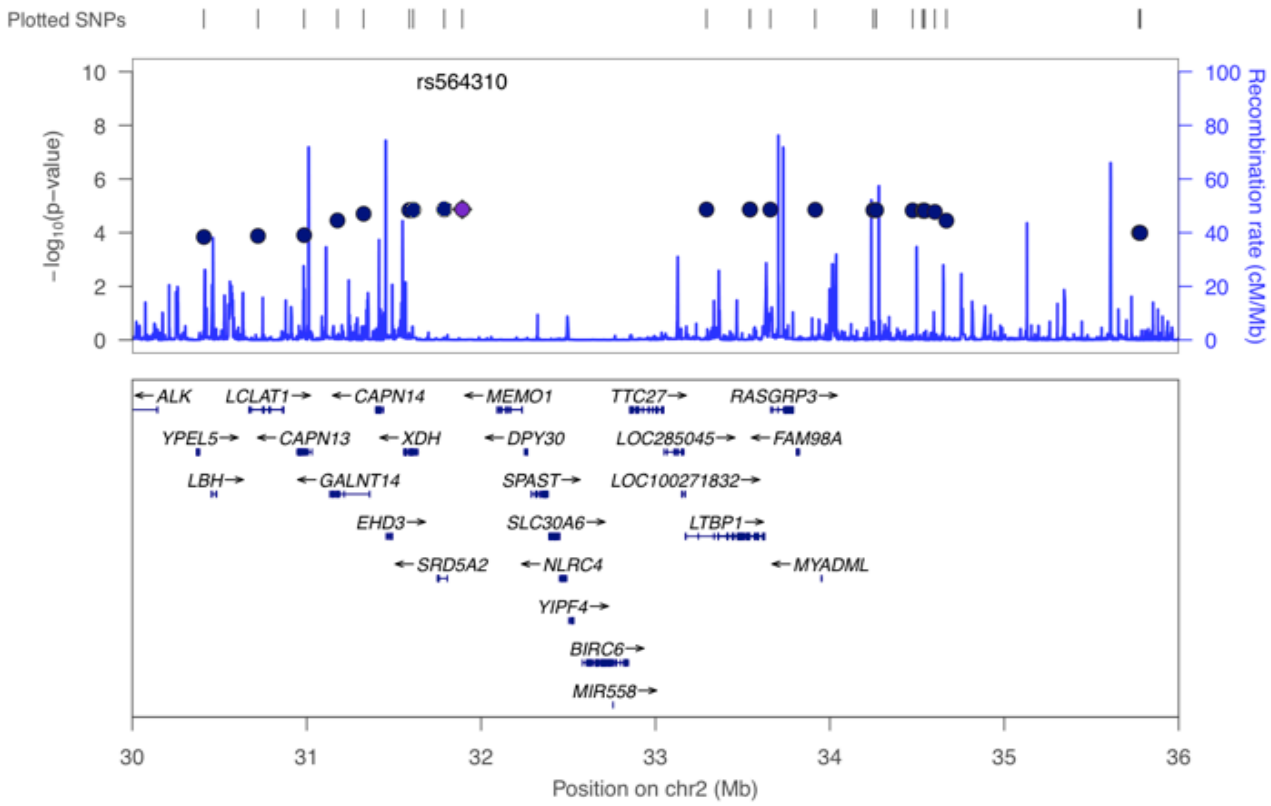


Figure 7.4. Graphical display of genes located on chromosome 2p23.1-p22.3 according to the human genome build 19. The graph depicts chromosomal positions in mega bases (Mb). The p-value for each SNP is plotted with the recombination rate between SNPs. P-values from the NPL_{pairs} analysis of Branch 1 under the super phenotype are presented in the graph. The purple diamond depicts the most significant SNP. Plotted with LocusZoom (Pruim et al., 2010).

As shown in Figure 7.4, the region on chromosome 2p23.1-p22.3 contains a number of interesting genes, including the EH domain-containing 3 gene *(EHD3)*, involved in endocytic trafficking. This was recently implicated with nominal significance (10×10^{-5}) in a whole-genome association study of 1022 cases with recurrent major depression and 1000 controls (Muglia et al., 2010). The identification of this gene in depression case-control cohorts further supports the potential relevance of this region in unipolar depression as well as BPD, as indicated by the findings from the BBF study. The region also includes the *ALK* gene, which encodes tyrosine kinase, a member of the insulin receptor super family that plays an important role in

neurodevelopment (Entrez Gene). Genetic variations of the *ALK* have been associated with schizophrenia in the Japanese population (Kunugi et al., 2006).

7.1.4 Chromosome 3p

The region on chromosome 3p24.3-p24.1 surpasses genome wide significant linkage in the BBF study (maximum LOD=4.18) under the narrow phenotype model with a dominant mode of disease transmission. With the exception of modest support to a marker on 3p24.3 (maximum LOD=1.19) reported in a study of 97 sibling pairs affected with BPI, BPII, SAD, and unipolar depression (Edenberg et al., 1997), this is not a previously implicated region in linkage studies of BPD, although there is evidence in both schizophrenia and depression if we consider the larger region on chromosome 3p25.1-p24.1 reported by MERLIN with near suggestive evidence for linkage.

A region on chromosome 3p26.2-p25.3, approximately 16.4 Mb more telomeric to the BBF near suggestive linkage peak, has been identified in a genome wide scan of 124 Indonesian sibling pair families with schizophrenia (LOD=3.76) (Schwab et al., 2008) and a region on chromosome 3p26-25, partially overlapping with the BBF linkage region, was reported with genome wide significant linkage (maximum LOD=4.0), in families with severe depression (Breen et al., 2010; Pergadia et al., 2011). These findings suggest that one region or multiple unrelated regions on chromosome 3p26-p24 confer susceptibility to BPD, severe depression and/or schizophrenia.

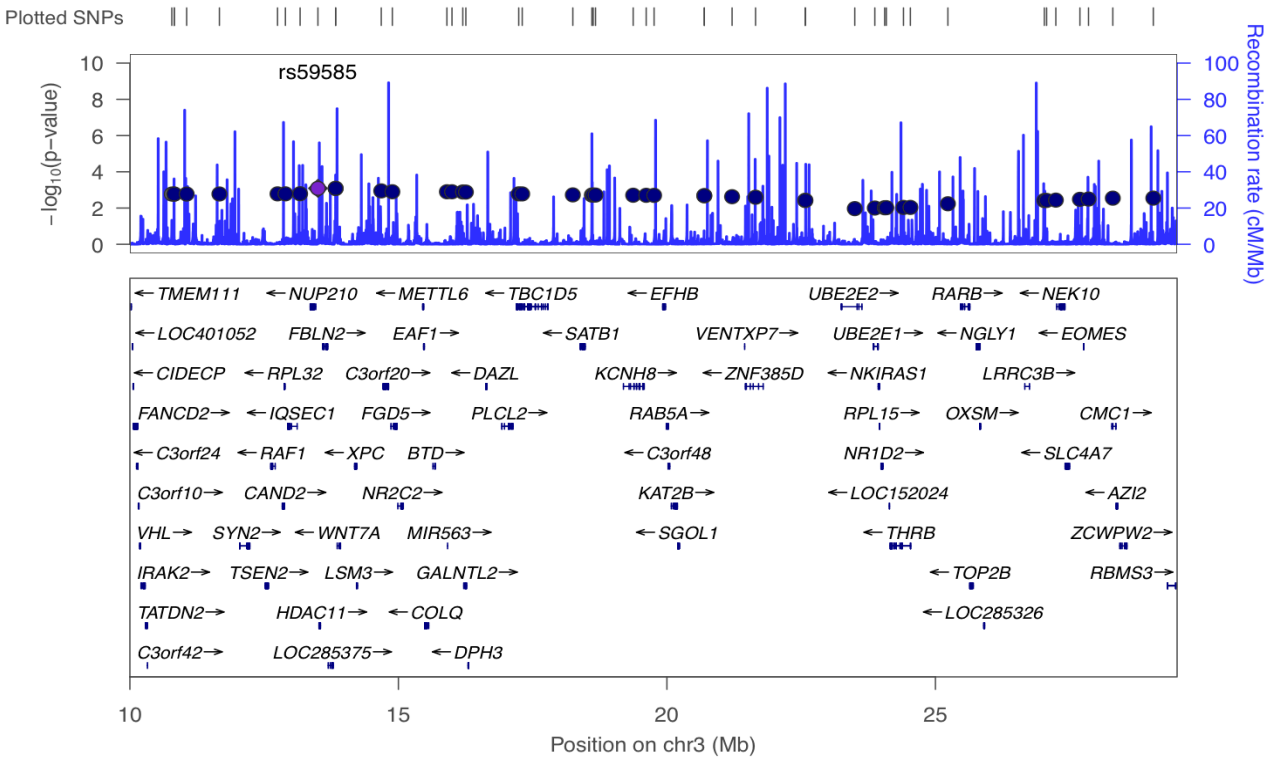


Figure 7.5. Graphical display of genes located on chromosome 3p25.3-p24.1 according to the human genome build 19. The graph depicts chromosomal positions in mega bases (Mb). The p-value for each SNP is plotted with the recombination rate between SNPs. P-values from the MERLIN analysis of Branch 1 under the narrow phenotype model are presented in the graph. The HLOD scores were converted to p-values by multiplying by 4.6 for a chi-square and then taking the value at 1 degree of freedom and dividing it by 2 (OTT,1999). The purple diamond depicts the most significant SNP. Plotted with LocusZoom (Pruim et al., 2010).

Of interest, in this region is the *NR1D2* gene (which is in the significant region implicated with significant linkage) involved in the regulation of circadian rhythms, although the extent of its involvement is yet to be determined (GeneCard). None of the other genes in the region have any relevance to BPD or mood disorders in general (Figure 7.5)

7.1.5 Chromosome 11p

The region on chromosome 11p15.4 achieved the second strongest genome wide significant linkage (maximum LOD=4.49) in the BBF under the depression phenotype model. Approximately 3.9 Mb telomeric to the BBF linkage peak, (Zubenko et al., 2003) reported a depression susceptibility locus on 11p15.5 with significant evidence for linkage (maximum LOD=4.20) using 81 families ascertained through probands with recurrent MDD. The findings from (Zubenko et al., 2003) are, however, difficult to interpret given that thirteen chromosomal regions revealed evidence of significant linkage in their study (Craddock & Forty, 2006). This region has also been implicated as a susceptibility locus for BPD (BPI, BPII and SAD) (maximum LOD=1.90[2]) in 153 multiplex bipolar families ascertained as part of the NIMH Genetic Initiative Bipolar Group on 11p15.5-p15.4 (Zandi et al., 2003). Their finding partially overlaps the BBF linkage region, however, it does not generalise to unipolar depression, which does differentiate it from the current study. This could be explained by the presence of genes that predispose to BPD in the region as well as genes that predispose to unipolar depression. Alternatively, this region could pleitropoic in that it confers risk to BPD in some families and to unipolar depression in others.

[2] This region was reported by Zandi et al. (2003) as an NPL score of 2.97 that was converted to a LOD score using LOD=NPL2/4.62 (Ott, 1999).

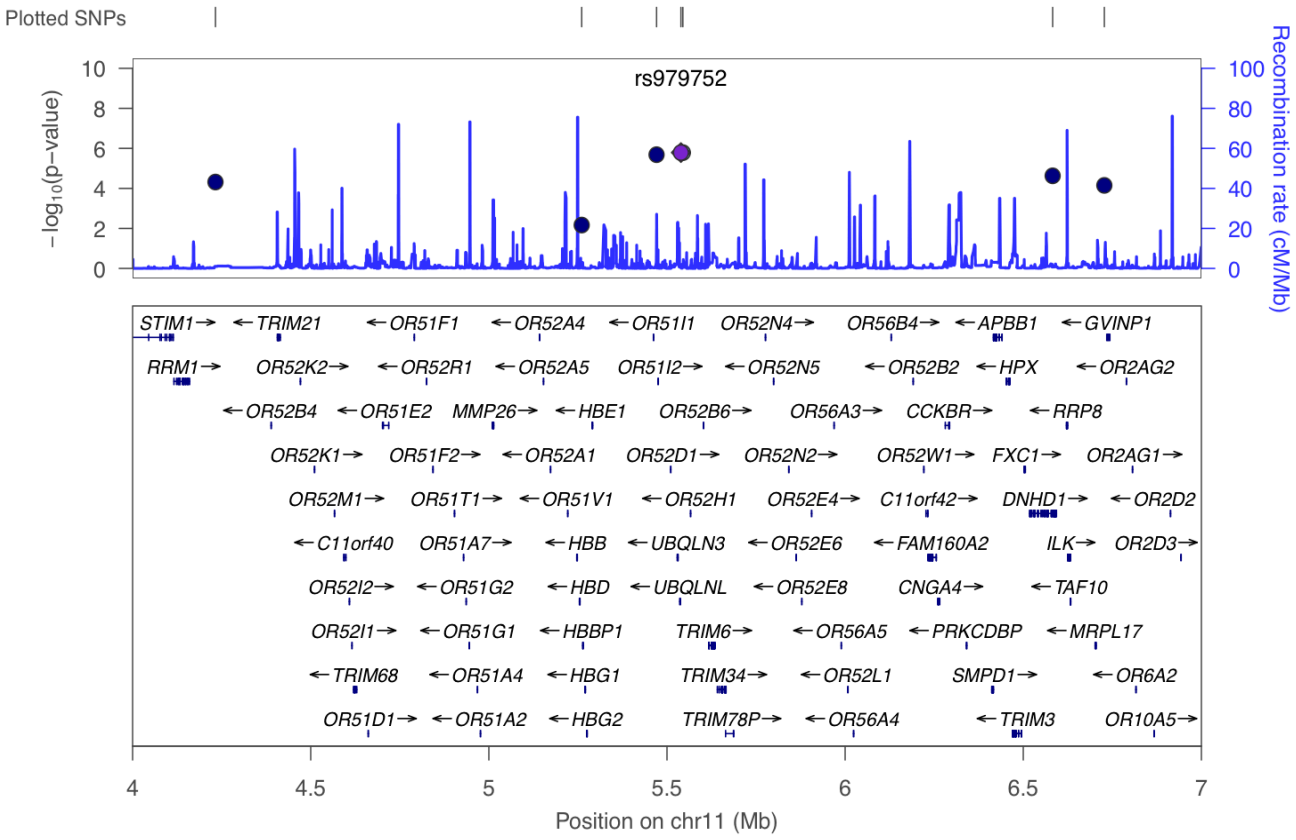


Figure 7.6. Graphical display of genes located on chromosome 11p15.4 according to the human genome build 19. The graph depicts chromosomal positions in mega bases (Mb). The p-value for each SNP is plotted with the recombination rate between SNPs. P-values from the NPL_{all} analysis of Branch 1 under the super phenotype are presented in the graph. The purple diamond depicts the most significant SNP. Plotted with LocusZoom (Pruim et al., 2010).

There are multiple genes of relevance in the region (Figure 7.6) including the *CCKBR* gene, also known as *CCK2*, which encodes the cholecystokinin B receptor protein (CCKB) and has a modulating effect on anxiety, analgesia, arousal, and neuroleptic activity (UniProtKB). CCKB activation appears to possess a general inhibitory action on dopamine activity in the brain, opposing the dopamine enhancing effects of CCKA (or CCK1). Several authors demonstrated CCKB receptor antagonists attenuate rodent anxiety (Griebel, Perrault, & Sanger, 1997), stop panic attacks in patients with a history of panic disorder (Bradwejn et al., 1994), and reverse pentagastrin-induced symptoms of anxiety in healthy

volunteers (Lines, Challenor, & Traub, 1995). More recently, the CCK neurotransmitter system has been implicated in the pathogenesis of panic disorder (Koefoed et al., 2010).

7.1.6 Chromosome 12p

Chromosome 12p13.32-p13.31 achieved genome wide suggestive evidence for linkage in the BBF (LOD=3.09) under the dominant mode of disease transmission. It has been previously implicated in a study of Columbian bipolar families, identifying a larger region that encompasses the linkage peak identified in this study on 12p11.21-q14.1 with suggestive evidence for linkage (maximum LOD=1.20[3]) (Kremeyer et al., 2010). Their finding was obtained using a broad phenotype model including BPI, BPII, and major depression, suggesting it as a mood disorder liability region. In contrast, the findings from the current study reported the effects of the region to be specific to depression. With the exception of the study by (Kremeyer et al., 2010) no linkage studies in mood disorders have previously implicated this region.

[3] This region was reported by Kryemer et al. (2010) as an NPL score of 2.35 that was converted to a LOD score using $LOD=NPL^2/4.62$ (Ott, 1999).

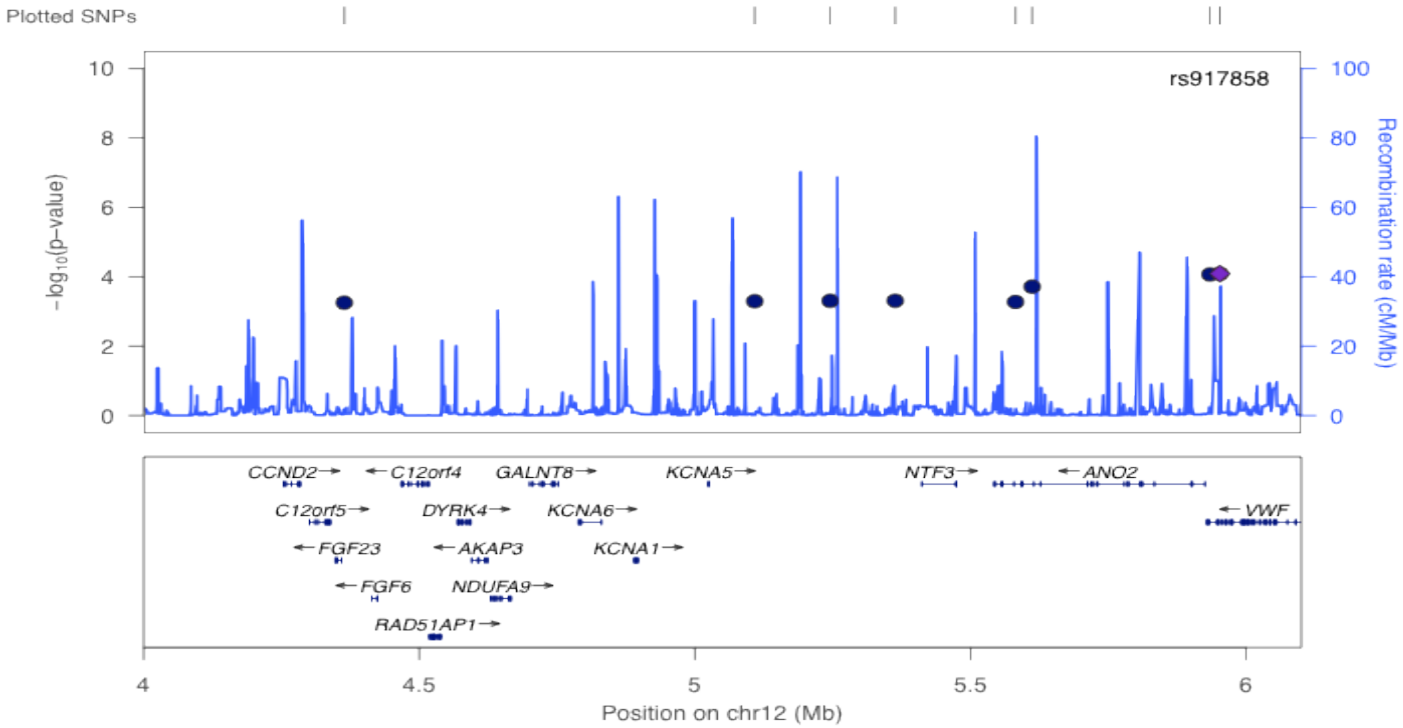


Figure 7.7. Graphical display of genes located on chromosome 12p13.32-p13.31according to the human genome build 19. The graph depicts chromosomal positions in mega bases (Mb). The p-value for each SNP is plotted with the recombination rate between SNPs. P-values from the McLinkage analysis of Branch 1 under the depression phenotype model are presented in the graph. The TLOD scores were converted to p-values by multiplying by 4.6 for a chi-square and then taking the value at 1 degree of freedom and dividing it by 2 (OTT,1999). The purple diamond depicts the most significant SNP. Plotted with LocusZoom (Pruim et al., 2010).

Despite not being identified as a susceptibility region for BPD, but depression specifically, it is still of interest that the BBF linkage peak lies approximately 3.7 Mb centromeric to a SNP (rs1006737) in *CACNA1C*, a confirmed gene in the aetiology BPD (Ferreira et al., 2008; Sklar et al., 2011). As indicated by Figure 7.7, the region on 12p13.32-p13.31 harbours a number of genes that could potentially prove relevant to the aetiology of mood disorders, for example, *NTF3*, which encodes a protein that is a member of the neurotrophin family, that controls survival and differentiation of neurons and is closely related to both nerve growth factor and brain-derived neurotrophic factor (GeneCard) and a family of voltage-gated potassium channel genes *(KCNA1, KCNA5,* and *KCNA6).*

7.1.7 Chromosome 12q

The highest LOD score reported in the BBF was on chromosome 12q24.22-q24.32 (maximum LOD=4.74) under the depression phenotype model. This region is supported by multiple genome wide significant signals in BPD and unipolar depression (Baron, 2002). Interest in the region on chromosome 12q23-24 (approximately 21 Mb centromeric to the BBF signal) started when it was described by a British pedigree co-segregating mood disorders and Darier's disease (maximum LOD=2.1), an autosomal dominantly inherited skin disorder (Craddock et al., 1994). Investigations of this region followed and a number of researchers implicated it in mood disorders (Dawson et al., 1995). Subsequently, Darier's disease was mapped to the *ATP2A2* gene, which encodes SERCA, a sarcoplasmic/endoplasmic reticulum calcium pump that plays a role in intracellular calcium signalling (Sakuntabhai et al., 1999) and researchers found the Darier-causing mutation did not play a major role in mood disorders (Jones et al., 2002). However, the presence of one or more BPD susceptibility genes in the Darier region has received some support from linkage studies (Green et al., 2005).

The BBF linkage region (12q24.22-q24.32) corresponds to a region implicated by a large pedigree from a homogeneous population in Québec (maximum LOD= 3.92) investigating individuals affected with BPD and unipolar depression (Morissette et al., 1999). In addition, in a region on 12q24.31, approximately 1.2 Mb cetromeric to the BBF linkage region, Shink et al., (2005) reported a susceptibility locus for mood disorders with highly significant LOD scores of 5.05 in BPD families from the isolated Saguenay-Lac-St-Jean population of Québec. The BBF region also overlaps with one identified by Ewald, Degn, Mors, & Kruse (1998) in two Danish families (maximum LOD= 3.37), however, unlike the findings of the current study, they reported evidence for linkage only when BPD patients were considered as affected. Further, approximately 27 Mb centromeric to the BBF

region Abkevich et al. (2003) identified a highly significant region on 12q22-12q23.2 (maximum HLOD=6.1) in males from 110 Utah pedigrees with a strong family history of major depression. The report that most closely mirrors the BBF finding was reported by Curtis et al. (2003) who found suggestive linkage (maximum LOD=2.8) on chromosome 12q24.31-q24.32 in seven families with multiple cases of BPD and unipolar depression. As with the BBF findings, they identified this region under a dominant mode of disease transmission and including unipolar cases as affected only.

Although, many of the regions discussed here do not match the specific one identified in the BBF, the localisation of susceptibility genes in complex diseases is difficult. Signals often are detected tens of centiMorgans away from the true locus (Roberts, MacLean, Neale, Eaves, & Kendler, 1999) with true linkage peaks usually being broader than false peaks (Terwilliger et al., 1997). Therefore, it is certainly plausible that the linkage peak identified in the BBF represents the same disease locus detected by the other studies described here. It is also plausible that the genes implicated in BPD and unipolar depression, in this region, are different or a collection of genes act in an additive or epistatic fashion to confer susceptibility to either BPD or depression. Further investigation of the region, possibly through sequencing may clarify its role in the different mood disorders.

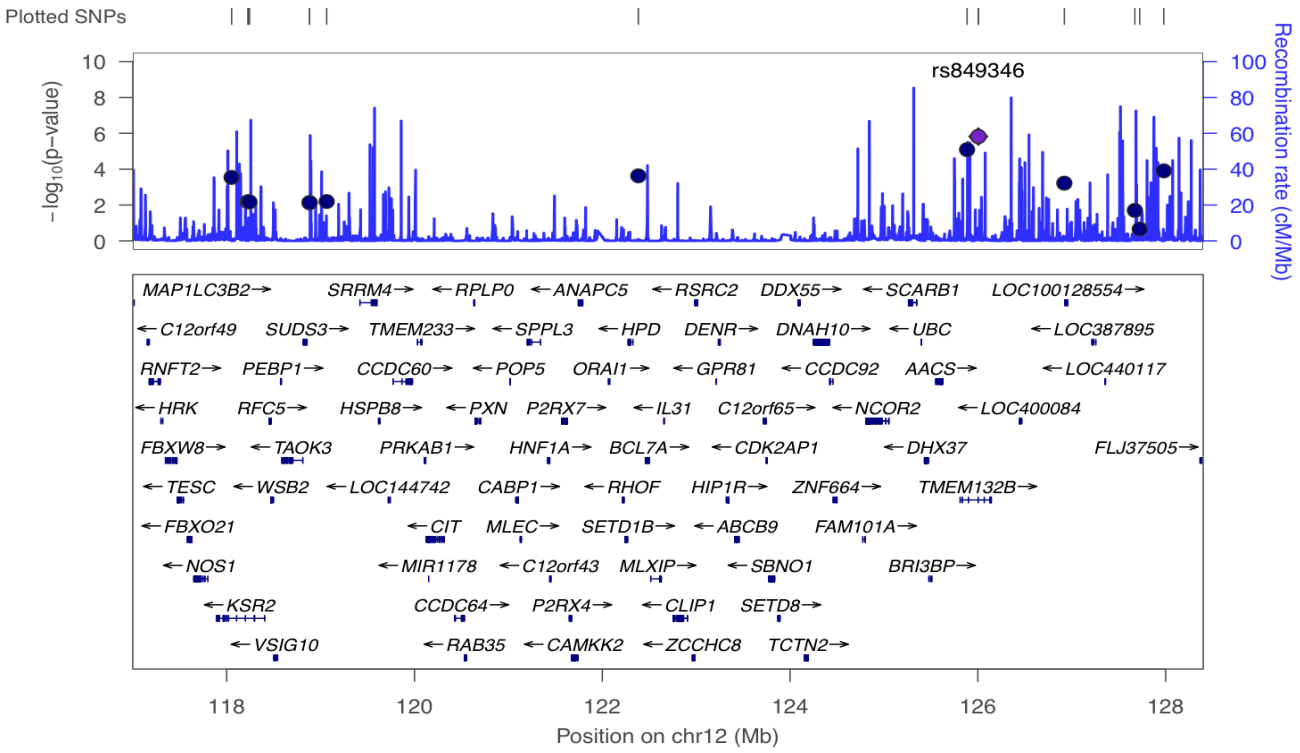


Figure 7.8. Graphical display of genes located on chromosome 12q24.22-q24.32 according to the human genome build 19. The graph depicts chromosomal positions in mega bases (Mb). The p-value for each SNP is plotted with the recombination rate between SNPs. P-values from the McLinkage analysis of Branch 1 under the depression phenotype model are presented in the graph. The TLOD scores were converted to p-values by multiplying by 4.6 for a chi-square and then taking the value at 1 degree of freedom and dividing it by 2 (OTT,1999). The purple diamond depicts the most significant SNP. Plotted with LocusZoom (Pruim et al., 2010).

Situated in this region are genes of interest (Figure 7.8) including *P2RX7*, previously associated with both BPD (Barden et al., 2006; McQuillin et al., 2009) and unipolar depression (Hejjas et al., 2009; Lucae et al., 2006). This association, however, was not replicated in two BPD studies (Green et al., 2009; Grigoroiu-Serbanescu et al., 2009). Another gene of interest in this region is the *NOS1* gene that belongs to the family of nitric oxide syntheses, and in the brain displays many properties of a neurotransmitter (Enterz Gene). It has been associated with Schizophrenia (Fallin et al., 2005) but does not show evidence for association with BPD in two studies (Buttenschön et al., 2004; Silberberg, Ben-Shachar, & Navon, 2010).

7.1.8 Chromosome 22q

Chromosome 22q11.21-q12.1 achieved genome wide suggestive evidence for linkage in the BBF (LOD=3.76) under a recessive mode of disease transmission and a broad phenotype model including BPI, BPII, SAD, BPNOS and cyclothymia. This region has been consistently implicated in psychiatric genetics in general. In BPD, Detera-Wadleigh et al. (1999) reported a suggestive linkage peaks that overlaps with the one identified in the BBF (LOD=2.1) in 22 U.S families affected with BPI, BPII, and SAD only. Edenberg et al. (1997) presented evidence for a linkage peak approximately 554 kb centromeric to the BBF peak in 540 individuals from 97 BPD families (NIMH bipolar initiative) with a maximum suggestive LOD score of 2.5 for affecteds with BPD and unipolar depression, thus implicating the region in general susceptibility to mood disorders. Kelsoe et al. (2001) reported a significant linkage peak 6.6 Mb telomeric to the BBF linkage peak in 20 U.S extended BPD families (maximum LOD=3.84) using a broad phenotype model that included BPI, BPII, SAD, and unipolar depression. The importance of this region in BPD and schizophrenia is further confirmed in a meta-analysis of eleven BPD linkage genome scans ($P < 1x10^{-5}$) and a meta-analysis of eighteen schizophrenia genome scans ($P < 9x10^{-5}$) (Badner & Gershon, 2002).

Interest 22q11 was initially driven by a reported association between BPD (and other psychoses) and velocardio-facial syndrome, a genetic disorder with a known microdeletion in 22q11 (Murphy, Jones, & Owen, 1999). Today there is mounting evidence both from quantitative and molecular genetic studies (e.g. Potash et al., 2003; Berrettini, 2003; Kohn & Lerer, 2005; Lewis et al., 2003) that suggest both BPD and schizophrenia share susceptibility genes in this region, offering support to the idea of a shared genetic aetiology for these disorders and suggesting that this region confers susceptibility to psychosis.

Chromosome 22q11.21-q12.1 was identified in the BBF under the broad phenotype model that included disorders that do not feature psychosis. Moreover, analysis under the narrow definition of affectedness, including the majority of cases with a history of psychosis showed no evidence for linkage. However, this might be accounted for by lack of power to detect linkage in the narrow group due to a smaller number of affected cases. Nonetheless the current study, in conjunction with the previous literature provides compelling evidence for the implication of this region in mood disorder as general liability.

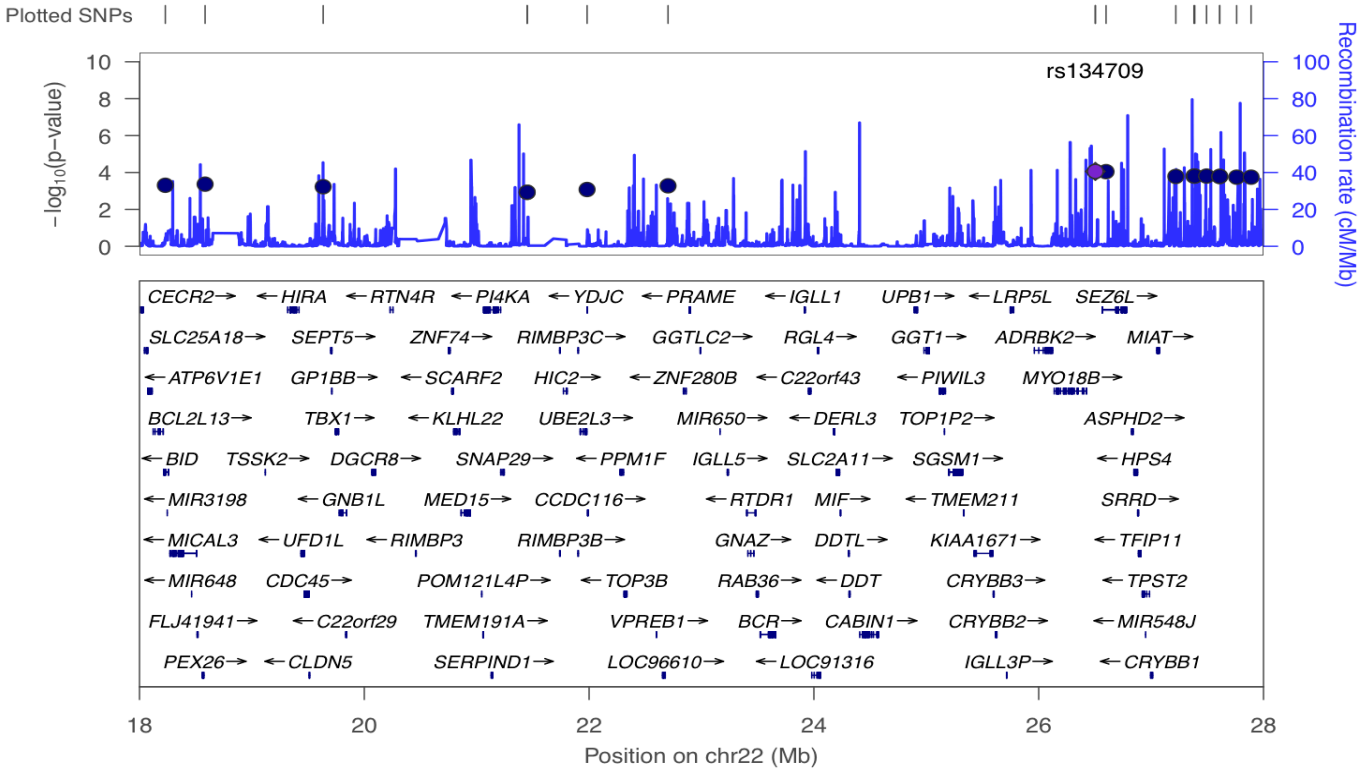


Figure 7.9. Graphical display of genes located on chromosome 22q11.21-q12.1 according to the human genome build 19. The graph depicts chromosomal positions in mega bases (Mb). The p-value for each SNP is plotted with the recombination rate between SNPs. P-values from the MERLIN analysis of Branch 1 under the broad phenotype model are presented in the graph. The HLOD scores were converted to p-values by multiplying by 4.6 for a chi-square and then taking the value at 1 degree of freedom and dividing it by 2 (OTT,1999). The purple diamond depicts the most significant SNP. Plotted with LocusZoom (Pruim et al., 2010).

There are multiple genes of interest within 22q11.21-q12.1 (Figure 7.9). These include some that are BPD candidate genes including *ADRBK2*, which encodes G-protein-coupled receptor kinase 3 *(GRK3)* involved in the phosphorylation of beta-adrenergic and related G protein-coupled receptors and has been associated with BPD (Barrett et al., 2003). Another gene of interest is the Catechol-o-methyl transferase *(COMT)* gene, which encodes enzymes involved in the degradation of monoamines. *COMT* contains the much studied Val108/158Met polymorphism which influences the enzyme's rate of activity and has been reported to influence clinical features BPD such as rapid cycling (Farmer, Elkin, & McGuffin, 2007). Other SNPs in the gene have also been reported in some studies to be associated with both BPD (Jones & Craddock, 2001; Rotondo et al., 2002) and schizophrenia (Palmatier et al., 2004; Shifman et al., 2002).

7.1.9 X-Chromosome

The X chromosome perhaps deserves special mention for historical reasons. It has recurrently been implicated since the very first attempt to discover genetic linkage in BPD with the observation of co-segregation of the disorder with colour blindness in one large pedigree (Reich, Clayton, & Winokur, 1969) and subsequently was the subject of several high profile early linkage studies with 'classical' makers (Baron, 2002). More recently there have been findings suggesting chromosome Xp22 and Xq26-28 (Stine et al., 1997) and Xq24-27 (Pekkarinen, Terwilliger, Bredbacka, Lönnqvist, & Peltonen, 1995) as susceptibility loci for BPD. None of these were replicated in the BBF study and it seems quite unlikely that X chromosome susceptibility loci have any role in BPD in this large BBF.

7.2 Interpretation of the Linkage Results

As discussed, the results of the whole genome linkage analysis conducted in this thesis support a number of regions previously identified in the literature. The question remains as to why identifying consistent and significant genetic susceptibility loci in linkage studies is difficult. A number of factors could be culpable including phenotypic heterogeneity, the oligogenic or polygenic nature of the disorder and the potential of genetic interactions between genes of small and large effect, individual genes having a small effect on the phenotype that are undetectable using linkage approaches, the influence of environmental factors, gene-environment interactions or correlations, and the influence of rare genetic variants. These factors will be discussed in the context of the findings linkage findings presented in this thesis.

7.2.1 Phenotypic Heterogeneity

The DSM-IV classifies psychiatric disorders into discrete entities with distinct boundaries. The adoption of this classification system into genetic studies was made under the assumption that the biological basis of these disorders was as distinct as the phenotypic manifestations. As previously discussed, BPD embodies a spectrum of symptoms and clinical characteristics that vary in kind and severity and whose underlying biological relationship is yet to be determined. Families with BPD segregate different subtypes of the disorder, e.g. BPI, and BPNOS, at different severity levels along with a variable number of co-morbid disorders such as anxiety disorders and alcohol related disorders. Further, the diagnostic boundaries differentiating BPD from depression are repeatedly challenged by family studies in which both BPD and unipolar depression co-segregate at higher rates than in the general population and from evidence that a large proportion of patients (40-60%) diagnosed with major depression either meet criteria for BPII disorder or subthreshold BPII

disorder characterised by hypomania that lasts for less than four days (Benazzi, 1999; Merikangas et al., 2007). These factors suggest that even though a growing literature supports the distinctness of BPD and unipolar depression at the genetic level (Breen et al., 2010), a general mood disorder liability is still plausible.

Identifying a genotype-phenotype relationship that describes the influence of a putative susceptibility locus on a disease is the essence of linkage mapping. The success of disease localisation depends to a large extent on how well the disease is defined, detected, or measured. The variability observed in linkage findings of BPD may, therefore, reflect the difficulty of defining the bipolar phenotype (MacQueen, Hajek, & Alda, 2005). To address the problem of phenotypic heterogeneity, using multiple phenotype models that vary the definition of affectedness under study is the standard in linkage studies of BPD. Most studies investigate a severe phenotype model including BPI and SAD, or BPI, BPII, and SAD, and a broad, less severe phenotype model that usually includes unipolar depression. This study included an additional model that included all of the subtypes of BPD to assess their genetic relatedness before considering unipolar depression. Four phenotype models; a narrow including BPI, BPII and SAD; a broad including narrow plus BPNOS and cyclothymia; a super including broad plus unipolar depression, and; a depression only phenotype model were analysed in the BBF linkage study.

7.2.2 Susceptibility to BPD, Unipolar Depression, and Mood Disorders in the BBF

Using hierarchal phenotype models does pose a challenge for the interpretation of the results, as increased evidence for linkage with a larger group of affecteds due to broadening the phenotype definition, could be attributed to an increase in power rather than to true genetic differences between the groups. By evaluating allele sharing statistics (available only from NPL tests) the interpretation of the results was facilitated. The BBF study presents some evidence for linkage under all investigated phenotype models. With the exception of chromosome 1p22.2-p21.3 being implicated with suggestive evidence for linkage under the narrow and broad phenotype models, none of the other chromosomal regions were implicated by more than one phenotype model.

7.2.2.1 Specific Liability to BPD in the BBF

Under the narrow phenotype model (BPI, BPII, SAD), locus 3p25.3-p24.1 yielded genome wide significant evidence for linkage. Broader phenotype models gave substantially lower LOD scores at this locus, indicating that susceptibility in this region is primarily to the more severe forms of BPD. This specificity could reflect underlying biological differences between the more severe forms of BPD, BPI, BPII, and SAD included in the narrow phenotype model, and the lesser severe mood disorders included in the other phenotype models.

Broadening the definition of affectedness to include BPNOS and cyclothymia implicated chromosome 22q11.21-q12.1 as a susceptibility locus for bipolar spectrum disorders. This implies that a disease causing mutation at this locus could result in variable symptoms, ranging from the more severe BPI disorder to the relatively mild cyclothymic disorder, in different affected individuals; it is pleiotropic. Why this region was not identified under the narrow phenotype model could be explained by lack of power to detect

linkage in the smaller number of affecteds in the narrow phenotype group or that the linkage signal is mainly driven by individuals affected with BPNOS and cyclothymia, as opposed to affecteds with more severe forms of BPD. However, it is important to address the fact that individuals classified under the broad phenotype model in the BBF are mainly children and adults whose mania was triggered by anti-depressant use. It is possible that these individuals will develop more severe forms of BPD later on and are arguably indistinguishable from individuals with BPI or BPII at the biological level. Association with depression, and therefore, more general liability to affective disorders, is dismissed at this locus based on the substantial decline in LOD scores observed with the inclusion of depression as an affected status in the analyses.

Interpreting the results from the region on chromosome 1p22.2-p21.3 was facilitated with the use of allele sharing statistics. Equal evidence for suggestive linkage in the narrow and broad phenotype models is contrasted by IBD allele sharing that is more than halved when the definition of affectedness is broadened to include BPNOS and cyclothymia. Ott (1999) explains this loss of linkage information could be attributed to misclassification of affecteds resulting in discordance between the phenotype and underlying genotype and therefore dilution of the linkage signals. Indeed this could be the case in the BBF study in which classification of individuals with BPNOS and cyclothymia as affected resulted in the observed reduction in allele sharing reported using the broad phenotype model. Accordingly, we could assume the linkage signal in 1p22.2-p21.3 is mainly driven by narrowly defined BPD.

The effort of this study to differentiate or examine the genetic liability of classic forms of BPD (BPI and BPII) versus BPNOS and cyclothymia often described as ‘soft’ BPD, yielded inconclusive results, probably due to the small number of affecteds with BPNOS and cyclothymia (n=12). Nevertheless, understanding the genetic relationship

between different forms of BPD is necessary for better treatment options in the future and is an area of research that warrants further investigation.

7.2.2.2 Specific Liability to Unipolar Depression

Findings from the study on chromosomes 11p15, 12p13.32-p13.31 and 12q24.22-q24.32 were specific to depression. The specificity of region 11p15 to depression is confirmed by extremely low LOD scores reported under the other phenotype models, as well as allele sharing scores that are approximately 10-fold lower in the narrow and broad phenotype models, whereas the super phenotype model, which includes depression, performs better than the two other groups in terms of LOD scores and allele sharing statistics. Similar patterns are observed in terms of LOD scores under the narrow, broad, and super phenotype models on the chromosome 12 regions. The argument concerning the dilution of the LOD score in the super phenotype model may apply here, where due to the misclassification of multiple non-affecteds (i.e. those with BPD diagnoses) as affected, the LOD score in the super phenotype model became non-significant. While unipolar depression and BPD may share susceptibility loci these three loci seem to only confer susceptibility to depression in the BBF.

7.2.2.3 General Liability to Mood Disorders

Susceptibility to mood disorders in general was implicated on chromosomes 1q21.1-q21.3 and 2p23.1-p22.3 under the super phenotype model, providing evidence they contribute susceptibility to both unipolar and BPD. These two chromosomal regions gave low, non-significant LOD scores when the narrow, broad, and, to a lesser extent, depression phenotype models were investigated. Evaluation of IBD sharing across phenotype models shows only a modest change in IBD sharing in the super model in comparison to the other

models, indicating linkage evidence was not realised in the other groups due to lack of power.

The presence of specific liabilities to both BPD and depression as well as a general liability to mood disorders is in keeping with twin study findings from McGuffin et al. (2003) who estimated the shared genetic influence in liability to mania and depression at approximately 29%, using twins ascertained from the Maudsley hospital twin register, indicating most genetic liability to mania (approximately 71%) is specific to bipolarity. Further, in line with the findings of this study are reports that unipolar depression is more common in the relatives of probands with BPD, while the opposite is not true (Jones et al., 2002; McGuffin and Katz, 1989; Weissman et al., 1984; Gershan et al., 1982), suggesting that families segregating BPD carry susceptibility loci for both BPD in particular and mood disorders in general, while families segregating unipolar depression are only enriched with genetic variants predisposing to depression or mood disorders in general but not BPD.

7.2.2.4 Limitation of Categorical Phenotype Definitions

A limitation of the linkage study concerns using categorical DSM criteria in defining disease phenotypes irrespective of clinical characteristics. Findings from linkage studies that use symptom dimensions and clinical characteristics have been more encouraging. For example, a linkage study on early onset BPD found a strong susceptibility locus on chromosome 9q34 for cases with mania onset before age 20 (Faraone, Lasky-Su, Glatt, Eerdewegh, & Tsuang, 2006). Another study attempting to address co-morbidity with BPD reported linkage on chromosome 6q25 with co-morbid suicidal behaviour, on chromosome 7q21 with co-morbid panic disorder, and on chromosome 16p12 for BPD with co-morbid psychosis (Cheng et al., 2006). Psychosis, as a measure of disease severity, was considered as a phenotype m9odel in the BBF linkage analyses, however, most of the individuals with psychosis were already considered under the narrow phenotype model and so conducting a

separate analysis with psychosis as a phenotype made little statistical sense.

In summary the linkage findings in the BBF family implicated chromosomes 1p22.2-p21.3, 3p25.3-p24.1, and 22q11.21-q12.1 in specific liability to BPD, chromosomes 11p12, 12p13.32-p13.31 and 12q24.22-q24.31 in specific liability to unipolar depression, and chromosomes 1q21.1-q21.3 and 2p23.1-p22.3 in a general liability to mood disorders.

7.3 The Mode of Bipolar Inheritance

The pattern of disease inheritance in the BBF does not support the presence of a founder effect where one or more major disease genes are inherited from a common ancestor and shared by all affected individuals in a pedigree. The whole genome linkage scan implicated eight chromosomal regions in the aetiology of BPD and unipolar depression in the BBF. Locus heterogeneity caused by "extraneous genes" introduced by marrying-in spouses is evident. Different combinations of the identified susceptibility loci were found to segregate in different subfamilies of the BBF (Table 7.1).

Phenotype Model	Susceptibility Loci	Branch1 Subfamilies (n=12)	Branches 2 and 3 (n=6)
Narrow	1p22.2-p21.3	1,2,3,5,7,9	None
	3p25.3-p24.1	2, 3, 8, 9, 10,12	14
Broad	1p22.2-p21.2	1,2,3,5,7,9,10	None
	22q11.21-q12.1	1,5,7,8,9,10	None
Super	1q21.1-q21.3	1,2,3,4,6,7,8,9,10,12	14,16
	2p23.1-p22.3	1,2,3,4,6,8,9,10,12	14
Depression	11p15.4	1,2,6,7,8,9,10,12	16

Table 7.1. Table showing the number of subfamilies positively contributing to the LOD score in the susceptibility loci identified using MERLIN. No single genetic locus imparted major risk for BPD or depression in all subfamilies in Branch 1 or the BBF. Although a subfamily's positive contribution to the LOD score does not necessarily mean that it is linked to the particular locus, this table gives an indication of the varied distribution of genetic susceptibility in the Brazilian subfamilies. Information on subfamily contribution to linkage regions 12p13.32-p13.31 and 12q24.22-q24.32 identified using McLinkage is not available.

Many other reports of densely affected large families and isolated populations where the segregation of major genes is plausible indicate locus heterogeneity and support the involvement of multiple genes in the aetiology of BPD (Morissette et al., 1999; Shink, Morissette, & Barden, 2003; Venken et al., 2005). In fact, with the exception of two segregation analyses in BPD suggesting the role of a single major gene in disease

transmission (Spence et al., 1995), the general consensus in linkage studies is that multiple genes confer susceptibility to BPD in nuclear and large families. The absence of a major susceptibility gene in BPD is also supported by mathematical modelling of the genetic transmission of BPD (Craddock, Khodel, Van Eerdewegh, & Reich, 1995).

The pattern of disease inheritance in the BBF is more appropriately explained in terms of a mixed model in which loci with medium to large effect sizes segregate in a dominant or recessive manner, against a background of polygenes with modest effect sizes that may modify or potentiate their effects. The effect sizes of the linkage regions were not tested, however, the assumption that they have medium to large effect sizes is supported by Risch & Merikangas, (1996) who estimated that loci conferring disease risk of two or less (i.e. small effects) are impossible to identify in linkage studies because more than approximately 2500 family members would be required to do so. A polygenic background of disease susceptibility is suggested by molecular genetic studies in sporadic cases of BPD that confirm the contribution of minor genes to BPD liability (Purcell et al., 2009; Sklar et al., 2011); however, their interactive effects remain unclear, both in association and linkage studies. It is possible that susceptibility to BPD and unipolar depression in the BBF is modulated by the interactive effects of different loci, including those identified in the study. To date, only two genome wide interaction linkage scans have been conducted to address epistasis in BPD. Both studies identified strong interactions between putative linkage loci suggesting epistatic effects contribute to the aetiology of BPD (Abou Jamra et al., 2007; Fullerton, Donald, Mitchell, & Schofield, 2010).

7.4 Environmental Factors

The whole genome linkage analysis conducted in this thesis did not account for the effects of the environment. Linkage methods were originally developed for Mendelian diseases that assumed the action of a single gene mutation was sufficient to cause disease. As such, environmental factors were not taken into consideration. This presents a methodological limitation to all linkage studies of complex multifactorial diseases.

We know that genes account for the majority of the risk to BPD, with heritability estimates in excess of 80%. However, this means that non-genetic factors account for approximately 20% of the variance to BPD in the population and are likely to have a role even in the special case of densely affected families. Twin studies show that it is non-familial environmental factors such as negative life events that influence susceptibility to BPD (Kendler, Pedersen, Neale, & Mathé, 1995; Kieseppa, Partonen, Haukka, Kaprio, & Jouko, 2004; McGuffin et al., 2003). Gene-environment interactions, the combined action of multiple genes of small effect together with a variety of environmental factors, have been shown to influence susceptibility to psychiatric diseases. In affective disorders the effects of gene-environment interactions have been limited to candidate gene studies. For example, fifteen independent studies have found an interaction between a common functional polymorphism in the promoter region of the serotonin transporter gene, *5-HTT*, and environmental adversity to be associated with the onset of depression (Uher & McGuffin, 2008). This has not been consistently supported in the literature, however studies that fail to present evidence for this particular interaction seem to test different environmental measures and use different methodologies (Karg, Burmeister, Shedden, & Sen, 2011; Uher et al., 2010). To date there is only one study investigating the impact of stressful life events on the propensity to develop depressive and manic episodes in BPD

patients; in this study depressive episodes were found to be moderated by a valine to methionine substitution polymorphism in the *BDNF* gene, but not manic episodes (Hosang et al., 2010).

7.5 Parametric and Nonparametric Linkage Results

7.5.1 Parametric Linkage Results

Parametric linkage analysis identified susceptibility loci on chromosomes 3p24.3-p24.1 (significant), 12p13.32-p13.31 (suggestive), 12q24.22-q24.32 (significant) and 22q11.21-q12.1 (suggestive) using McLinkage analyses conducted on Branch 1 and the BBF as connected family unit under the assumption of genetic homogeneity. Evidence for linkage was only attained on chromosome 22q11.21-q12.1 (suggestive) using MERLIN analyses on subfamilies from Branch 1 and the BBF under a model of heterogeneity. The region on chromosome 22q11.21-q12.1 was the only region to receive support from both linkage programs and MERLIN achieved higher HLOD scores than McLinkage indicating a level of heterogeneity, with the subfamilies in Branch 1 yielding a maximum LOD score of 3.76 and the unit of Branch 1 yielding a maximum LOD score of 3.53.

Chromosome 3p24.3-p24.1 achieved genome wide significant linkage (maximum LOD=4.18) in Branch 1 and the BBF using McLinkage and a larger region on chromosome 3p25.1-p24.1 achieved near suggestive evidence using MERLIN. One could deduce that this is a true linkage peak that failed to reach significance using MERLIN due to loss of power from splitting the pedigree into subfamilies. However, one needs to keep in mind that the region only achieved significance in McLinkage at a recombination fraction of 0.2 indicating that the disease locus is approximately 2 cM away form the marker locus. Fine mapping studies will clarify the role of this region in BPD.

It is perplexing as to why the region achieving the highest LOD score in the study on chromosome 12q24.22-q24.32 using McLinkage was not detected using MERLIN. In fact HLOD scores reported in this region by MERLIN equalled or were slightly greater than zero. This region could be a false positive given the small size of the linkage region (120 kb) and the reported sharp decline in LOD scores. However this region is commonly associated with mood disorders in general, so it warrants some consideration. Having said that and as will be discussed in *section 7.6* the McLinkage analyses were somewhat problematic and therefore the results need to be considered with some caution.

7.5.2 Nonparametric Linkage Results

Non-parametric linkage analysis identified susceptibility loci on chromosomes 1p22.2-p21.3 (suggestive), 1q21.1-q21.3 (suggestive), 2p23.1-p22.3 (significant), and 11p15.4 (significant) using MERLIN. It is important to interpret these results in light of the properties of the NPL statistics used to identify them. By estimating IBD between pairs of affected individuals, NPL_{pairs} may attribute significance to regions where multiple variants confer susceptibility to disease in different pairs of affected family members. Conversely, due to estimating of IBD scores across groups of affecteds, NPL_{all} may only attribute significance to regions where a single variant confers susceptibility to disease (Kong and Cox, 1997). Therefore, NPL_{pairs} allows for allelic heterogeneity within family members in each subfamily, while NPL_{all} does not. Interpreting the findings of the NPL with respect to this allows us to draw some conclusions. Significant NPL was achieved on chromosome 11p15.4 using NPL_{all} and on chromosome 2p23.1-p22.3 using NPL_{pairs} only, while suggestive NPL was achieved on chromosome 1p22.2-p21.3 using NPL_{all} and on chromosome 1q21.1-q21.3 using NPL_{pairs} only. Given the properties of these test, we can deduce that the same genetic variants on chromosomes 11p15.4 and 1p22.2-p21.3 confer

susceptibility to disease within and perhaps between the sub-subfamilies, while different and perhaps multiple variants within chromosomes 1q21.1-q21.3 and 2p23.1-p22.3 confer susceptibility to disease in the subfamilies.

7.5.3 Parametric Versus Nonparametric Linkage Results

Parametric and NPL methods yielded different results in the linkage analyses. Only two regions were corroborated with LOD scores greater or equal to one across linkage methodologies, chromosome 11p15.4 and 22q11.21-q12.1. One could reason the susceptibility loci identified using parametric linkage had segregation patterns compatible, or near compatible, with the specified genetic models and were therefore identified using parametric linkage. They were potentially not identified by NPL due to lack of power. Equally, susceptibility loci identified using NPL possessed more complex patterns of inheritance than the models specified under parametric linkage and were therefore undetected. Furthermore, unaffected individuals many of whom presumably carry disease alleles without expression of disease are not considered in NPL methods. With parametric linkage one can weight evidence of linkage that some unaffected individuals harbour the disease allele, but not express illness by using penetrance functions. However, they are difficult to estimate and could account for discrepancies between the two linkage approaches.

7.6 Large Family Studies

It has often been proposed that large families are ideal samples for the molecular investigation of BPD susceptibility genes because they may segregate highly penetrant variants that predispose to disease, thus making them easier to detect. Highly penetrant variants also mean that affected family members are less likely to represent phenocopies. In addition, information on linkage phase, anticipation, and imprinting is more readily available in large families, which enables more accurate definitions of the mode of disease transmission and penetrance parameters necessary for parametric linkage (Blackwood et al., 2001). One large multigenerational family might also have the advantage of a more consistent environment over a collection of multiple nuclear families.

The benefit of analysing a large family as a connected unit were somewhat realised in the analysis of the BBF with the identification of two depression susceptibility loci on chromosomes 12p13.32-p13.31, and 12q24.22-q24.32 that would have otherwise been missed had the analyses been limited to those conducted on the subfamilies. However, the consistency of the findings generated using McLinkage where difficult to determine. Initially the Marcov Chain Monte Carlo (MCMC) based method *(see chapter 4 section 4.5.4.6)* was run using 1,000 iterations that yielded inconsistent LOD scores every time the analyses were run, indicating a lack of convergence. Subsequent analyses were performed using 10,000 iterations, which posed a number of computational difficulties and took approximately four months to complete. This long run time is partially due to the large number of closely spaced SNPs in the BBF genome scan, as opposed to the couple of hundred microsatellites usually employed in linkage scans of bipolar pedigrees, which exponentially increases the time required to achieve convergence. Consequently, the stability of the McLinkage results (all of which were the product of 10,000 iterations) presented in this thesis was not assessed. The differences in linkage findings between

Branch 1 and the BBF (which appear to follow a similar pattern in MERLIN with Branch 1 yielding more linkage evidence) as well as the observed sharp oscillations in LOD scores could be due to the non-convergence of the program.

The increased power to detect linkage in a large connected multigenerational pedigree was demonstrated in a microsatellite based study (459 microsatellite markers) that successfully analysed a thirteen-generation pedigree from the Central Valley of Costa Rica that was previously analysed as disconnected units. Using Simwalk2 (Sobel & Lange, 1996), a linkage program based on the MCMC method used in McLinkage, estimation of allele sharing among affected relative pairs confirmed previously reported linkage to chromosome 18 and identified suggestive evidence for linkage on a new locus on chromosome 5q31.3-33.2 (Garner et al., 2001), which was subsequently confirmed by fine mapping studies (Herzberg et al. 2006; Jasinska et al. 2009).

Nevertheless, analysing large pedigrees using dense SNP maps, from the experience of the BBF study, and as indicated by reports in the literature is fraught with limitations. Similar computational difficulties to those experienced in analysing the BBF were reported by Service et al. (2006) who attempted to analyse SNP data (n=4,690) in a bipolar pedigree from the Central Valley of Costa Rica comprised of 168 members, 82 of whom were genotyped, and found the pedigree in its entirety was too large and complex for multipoint analysis using any available software. Attempts to use Simwalk2 were unsuccessful due to the program's inability to converge after weeks of running. A similar scenario was encountered by Marcheco-Teruel et al. (2006) who attempted to conduct a genome wide scan of a highly consanguineous six-generation pedigree from a relatively isolated region in eastern Cuba, with 73 family members genotyped using the Affymetrix 10K SNP marker set (n=10043). The large size and complexity of the pedigree entailed subdividing it into four branches for the analyses.

The difficulties of linkage conducted using MCMC based approaches could be avoided with the use of exact LOD score calculations. However, this method is unfeasible with large complex families unless they are split into smaller units, which results in loss of inheritance information. At this point in time, taking advantage of the genetic information in large pedigrees is hindered by inadequate analytic tools.

7.6.1 Inbreeding Loops and Increased Homozygosity

The presence of inbreeding loops in the BBF posed a second challenge to the linkage analyses. Most linkage programs rely on the process of clipping or peeling, in which small nuclear families within a large pedigree are analysed and all the information is collapsed onto one of the parents or relatives whose own sibship is analysed next. This continues until all the information is collapsed into one final person, the founder. If a loop exists in a pedigree, the process of collapsing circles around and around in an infinite loop (Terwilliger & Ott, 1994a). Inbreeding also increases the frequency of homozygosity, which decreases the informativeness of marker sets, subsequently reducing power to detect linkage. This is because marker heterozygosity is an important factor that helps determine whether a parental gamete is recombinant or non-recombinant, necessary for parametric LOD score calculations, and to differentiate between IBD and IBS necessary for NPL calculations (Ott & Rabinowitz, 1997). The Affymetrix 10K marker set used in the analyses captured 81% and 65% of the information content of Branch 1 and the BBF respectively. The suboptimal coverage of the genome, which may have lead to significant linkage being missed, is possibly a consequence of increased homozygosity in the family but is also likely to be a consequence of splitting the BBF into small subfamilies.

Homozygosity mapping is based on the assumption that homozygous segments are inherited from an ancestor common to the maternal and paternal lineage without recombination and has been effective in detecting genetic mutations in rare recessive

diseases. Its applicability to complex disease has been limited to identifying modifier genes inherited in a recessive manner in conjunction with major disease loci (Ewald, Kruse, & Mors, 2003). The description of homozygosity mapping in the literature is confusing and warrants explanation. Some researchers refer to linkage analysis using a recessive mode of disease transmission as homozygosity mapping (Ewald et al., 2003; Ewald et al., 2005). This I have completed as part of the parametric linkage investigation conducted in this thesis with only a region on chromosome 22q11.21-q12.1 yielding suggestive evidence for linkage to disorders defined under the broad phenotype model (BPI, BPII, SAD, BPNOS, cyclothymia). However, more recent approaches to homozygosity mapping, e.g. PLINK Runs of Homozygosity and IBDFinder (Carr, Sheridan, Hayward, Markham, & Bonthron, 2009) involve scanning the genome for homozygous segments and evaluating their association with diseases in a case-control comparison between affected and unaffected family members. These methods require high density SNP maps, and therefore were inapplicable to the BBF study.

7.7 Anticipation

Anticipation, characterised by earlier disease onset in successive generations was reported in the BBF *(see chapter 2 section 2.1.5).* In line with the findings in the BBF, McInnis et al., (1993) found members of the younger generations in 34 unilineal families with BPD manifested significantly more severe forms of disease on the basis of age of onset and episode frequency. Genetically, anticipation correlates with the expansion of trinucleotide repeat sequences. After reaching a critical size, they show a tendency to increase in size during both meiosis and mitosis. They are known to cause disease when located near or adjacent to specific genes. A method called repeat expansion detection (RED) was

developed to allow the detection of large trinucleotide repeats in human genomic DNA (Schalling, Hudson, Buetow, & Housman, 1993). To date, at least nine disorders including fragile X (Mahadevan et al., 1992), myotonic dystrophy (Fu et al., 1991), and Huntington's disease (MacDonald et al., 1993) have been found to be associated with these dynamic mutations.

There are suggestions that genes with expanding trinucleotide repeat sequences may have a role in the genetic aetiology of BPD (and schizophrenia). Three RED studies showed that the length of the most common pathogenic trinucleotide repeat, CAG/CTG, was greater in individuals with BPD and schizophrenia than in healthy controls. These findings were replicated in a large multicentre European consortium; however evidence that maximum repeat size was related to age of onset was not found (O'Donovan & Owen, 1996). Some subsequent studies partially replicated the findings (Del-Favero et al., 2002; Lindblad et al., 1998), while others failed to do so (O'Donovan, Jones, & Craddock, 2003). Further, the evidence got less convincing when studies found that genes in the vicinity of the two loci where the majority of the trinucleotide repeat expansions involved in BPD and schizophrenia occurred (90% of repeat trinucleotides detected by RED in BPD and schizophrenia occurred on chromosomes 18q21.1 and 17q21.3) were not associated with either BPD or schizophrenia (Vincent, Paterson, Strong, Petronis, & Kennedy, 2000). The lack of association with adjacent genes does not, however, mean that this trinucleotide repeat is not important. The trinucleotide repeat might have an impact on the structural, epigenetic, and/or translational aspects of genes – independent of those neighbouring genotypes. Today, there is no consensus as to whether trinucleotide repeats contribute to the aetiology of BPD or schizophrenia. Caution in interpretation of the reported anticipation in the BBF is advisable until further investigation as the anticipation in the family could be explained by environmental factors and/or an ascertainment bias.

7.8 Case Control Replication Study

In chapter 6, the contribution of the gene or genes driving the linkage signals in the BBF to disease susceptibility in sporadic BPD and depression cases was tested for association with disease in two large bipolar and depression case control cohorts. No associations with BPD or depression were found. This could be explained by a number of factors. Firstly, the susceptibility loci identified in the BBF may be a result of multiple rare variants that do not play a role in the aetiology of most cases of BPD and unipolar depression. Secondly, the linkage signals in the BBF may reflect the effects of multiple rare variants that are impossible to tag with the available SNP genotyping technologies that are based on common polymorphisms. Finally, the susceptibility loci identified in the BBF could be a result of combinations of common genetic variants, however their effect sizes are small in the general population and can only be detected with sample sizes larger than the two case-control cohorts used in this thesis. It is most likely that the linkage peaks identified in the BBF reflect a combination of rare and common genetic variants that contribute to disease susceptibility in the BBF and a combination of the mentioned factors prevented their identification in the case-control cohorts. The presence of common and rare disease causing variants in complex diseases is supported by findings from diabetes, autoimmune disorders, cardiovascular disease and Alzheimer's Disease (Badner et al., 2011).

7.8.1 The Generalisability of Findings from Large Family Studies

The common belief that disease causing genes identified in large families are rare, and therefore account for a small proportion of liability to disease in the larger population, is not entirely justified. Linkage analyses of complex diseases have enhanced the understanding of the molecular pathophysiology of a number of diseases, and have helped identify common disease causing variants that confer susceptibility to non-familial forms of

diseases, as in Alzheimer's disease (AD). The localisation and identification of the first AD locus, the gene that encodes the Amyloid Precursor Protein *(APP)* lead to the formulation of the Amyloid cascade hypothesis that explained the aetiology of AD (Hardy & Higgins, 1992). This hypothesis has been since corroborated by the subsequent identification of additional loci *PSEN1* (Sherrington et al., 1995), *PSEN2* (Levy-Lahad et al., 1995; Rogaev et al., 1995), and *APOE* (Pericak-Vance et al., 1991); all thought to act together in one pathway. Although familial forms of AD constitute a minority of all cases, the *APP*, *PSEN1* and *PSEN2* gene were all identified through linkage analysis in extended pedigrees. Subsequent findings confirmed the role of APOE4 in late onset AD that is more common in the general population by association analysis (Payami, Kaye, Heston, Bird, & Schellenberg, 1993). While it is uncertain whether this example could be transferred to complex psychiatric diseases that do not appear to have Mendelian forms, it shows that the identification of loci that have a direct impact on a small subset of individuals might provide an essential step towards the understanding of the molecular mechanisms of disease. In psychiatry, Disrupted in Schizophrenia *(DISC1)* is an example, first found to be disrupted by a balanced 1 to 11 translocation segregating with severe psychopathology, including schizophrenia, BPD, and mental retardation, in one extended pedigree from Scotland (St Clair et al., 1990). The research group subsequently cloned the break point on chromosome 1q42 and identified two genes they called *DISC1* and *DISC2* (Millar et al., 2000). Association with common SNPs within *DISC1* have since been found in unrelated individuals with schizophrenia and BPD in the general population (Hennah et al., 2003; Hodgkinson et al., 2004; Schosser et al., 2010). However, the findings have been inconsistent and are likely due to discrepancies in samples sizes, clinical phenotypes, and study methodologies (Chubb, Bradshaw, Soares, Porteous, & Millar, 2007; Prata et al., 2010).

7.8.2 Common Genetic Variants in Complex Disease

Linkage studies lost some favour in human genetics following a host of inconsistent and un-replicable findings. The advent of high throughput genotyping and possibility of analysing case-control data using a hypothesis-free approach saw a surge in genome wide association studies (GWAS). Today, GWAS have been completed for most psychiatric diseases and have probably identified most of the common genetic variants (minor allele frequencies greater than 5%) involved. However, most of the associated SNPs have very small effect sizes and the proportion of heritability explained is at best modest for most psychiatric diseases. The term missing heritability refers to the discrepancy between total heritability as estimated by twin and family data, and the proportion of phenotypic variation explained by all detected SNPs from GWAS that are associated with the disease at genome wide significance levels (Visscher, Goddard, Derks, & Wray, 2011). A recent collaborative BPD GWAS including a total of 4,387 cases and 6,209 controls identified only two regions of genome wide significant association in *ANK3* (9.1×10^{-9}) and *CACNA1C* ($P=7.0 \times 10^{-8}$) each with an odds ratio below 1.45 despite a heritability of approximately 70%-90% for BPD (Sklar et al., 2011). Further, hits from GWAS, or SNPs in LD with hits from GWAS, have rarely been tracked to causal polymorphisms and give no insight into disease mechanisms, leading many to assume that the associated variants must have subtle regulatory effects (Cirulli & Goldstein, 2010) (i.e. control the expression of one or more other genes. A regulator gene may encode a protein, or it may work at the level of RNA, as in the case of genes encoding microRNAs.).

These observations from GWAS lead to suggestions that rare genetic variants defined by allele frequencies of less than 1% may contribute to common diseases (Bodmer & Bonilla, 2008). An increasing body of evidence from cardiovascular research and other complex diseases, indicates rare genetic variants explain a significant amount of disease

heritability. A series of studies that have re-sequenced genes of known function in individuals with cardiovascular disease have identified rare coding variants with substantial effects on levels of cholesterol, triglycerides, etc (Romeo et al., 2009). In addition, re-sequencing of genes identified in GWAS for hypertriglyceridemia has concluded that there is an excess of rare variants from these genes in cases compared to controls (Johansen et al., 2010). The current suggestion is that GWAS signals may reflect the effect of multiple rare variants that have been credited to common variants, so called synthetic associations and sequencing or even targeted re-sequencing of candidate genes highlighted from GWAS data may reveal enrichment of rare variants in cases versus controls (Dickson, Wang, Krantz, Hakonarson, & Goldstein, 2010).

Reassessment of the contribution of rare genetic variants to disease susceptibility renewed interest in linkage studies particularly those conducted in large families and isolated populations. Large families facilitate the detection of rare disease causing variants because they are present at much higher frequency in affected family members. Families with multiple early-onset cases are likely to be enriched with variants of large effect as well as variants of small effect. Isolated populations, such as the Icelandic population, are used to localise disease predisposing genes by tracing unexpectedly long runs of identity-by-descent sharing among affected relatives (Kong et al., 2008). In retrospect, inconsistent linkage findings attributed to locus heterogeneity, reduced penetrance and phenocopy rates, terms that pertain to Mendelian diseases, could be explained in terms of multiple undetectable rare genetic variants (in linkage due to small effect sizes) in the aetiology of the disease (Cirulli & Goldstein, 2010).

7.8.3 Whole Genome Sequencing

Whole genome sequencing offers a comprehensive collection of rare variants and structural variation (deletions, duplications, copy-number variants, insertions, inversions and

translocations) for study. DNA sequencing technology has been available for many years; however recent advances in DNA sequencing technology have vastly increased the volume of sequence that can be obtained in a short period of time. New sequencing machines can currently sequence 25 billion base pairs of sequence a day (HiSeq 2000 from Illumina), although this rate has been increasing with remarkable rapidity. Large amounts of data provide sufficient coverage to identify most variants present in the genome.

Large scale whole genome sequencing projects are currently under way. The 1000 Genomes Project is an international research collaboration established to undertake sequencing of the complete genome of approximately 2500 individuals. The first part of the study sequenced two mother-father-adult-child trios from the HapMap project at high coverage, with each single base sequenced 20 to 60 times, and 179 individuals at low coverage, with each single base sequenced 2 to 4 times. They also sequenced 1,000 genes in 900 individuals at high coverage (50X). In the project's second phase, the genomes of 2,500 people will be sequenced at a rate of more than two genomes every 24 hours. The project aims to be able to identify very rare variants with a MAF of 0.5% and greater.

Sequencing families with multiple affected members has been recommended (Manolio et al., 2009). Sequencing nuclear families enables the detection of approximately 70% of sequencing errors and permits the identification of precise locations of recombination events. This results in turn to near complete knowledge of inheritance states by precisely determining the parental chromosomal origins of sequence blocks in offspring (Roach et al., 2010). Alternatively, sequencing family members with available linkage data is useful in prioritising variants identified in whole genome sequencing. It is a very efficient approach to sequencing that takes advantage of existing family data that is difficult to collect. Families driving the linkage evidence are re-contacted (granted consent is available) and enrolled in follow-up sequencing projects (Bowden et al., 2010).

7.9 Future Directions

7.9.1 Sequencing the BBF

It is the objective of the BBF study to identify functional or causal mutations within the identified linkage regions that may contribute to the development of BPD and depression using next generation sequencing approaches. The first stage of the sequencing analysis of six BBF members has already begun. The following methodologies were used. To qualify for this sequencing study family members needed to have a BPI disorder diagnosis with psychotic features, and to belong to a subfamily showing evidence of linkage to one or more of the regions identified with whole genome significance in the BBF linkage analyses. Family members were then selected if they had PLINK Identity-by-Descent (IBD) pair-wise estimates (--genome) (*see Chapter 4 section 4.4.2)* greater than 0.25 (i.e. second degree relatives), as the more distantly related the affected individuals the fewer genetic variants they will share. Whole genome sequencing was outsourced to Beijing Genomics Institute (BGI). Following receipt of the sequence data from BGI, data cleaning and alignments to the reference genome were conducted. Variant calling will be completed in the near future.

For a genetic variant (e.g. SNPs, CNVs) to truly have an effect on disease aetiology, it should somehow perturb either the protein itself or gene expression. Sequence analysis of the BBF will focus on variants with functional consequence for protein-coding genes in the linkage regions. For example variants that delete some or all of a gene, introduce a premature stop site into a protein or result in a non-conservative amino acid substitution. The aim is to identify genetic variants with functional effects and elucidate their biological mechanism in BPD.

7.9.2 Collaborative Studies

The complex nature of BPD genetics has meant that the identification of susceptibility loci has remained difficult. To this end, the future of mood disorders and psychiatric genetics in general may be in large international collaborations, involving thousands or even tens of thousands of samples. For example, collaborations in GWAS offer increased sample size, thereby also increasing the power to detect associations of small effect. The psychiatric GWAS consortium (PGC) is one such mega collaboration that has recently reported on a combined GWAS of 7481 individuals with BPD and 9250 controls (Sklar et al. 2011).

Common variants, however, are not solely responsible for the genetic aetiology of mood disorders. The percentage of individuals with mood disorders who may harbour rare genetic variants responsible for the manifestation of the disorder is currently unknown, but future re-sequencing projects will provide answers. In order to reduce type I error, re-sequencing projects will have be performed in large cohorts of cases and controls as well as families to accurately define private (i.e. occurring in a single family) versus rare mutations. The future of BPD genetics is in large scale collaborative projects which are necessary for the sufficient sample sizes to be able to detect genetic variation of both small and large effect. However, the necessity for large sample sizes in BPD should not overshadow the need for stringent inclusion criteria that clearly define the disorder in terms of severity as well as explore its relatedness to unipolar depression and schizophrenia.

7.10 Conclusion

This thesis investigated the aetiology of BPD in a complex, multigenerational family from Brazil, with the aim of contributing to the current understanding of the genetic determinants of BPD. The BBF study found four genome wide significant regions that conferred susceptibility to BPD, unipolar depression, and mood disorders in general confirming the presence of a specific and a shared liability to BPD and unipolar depression. The significant linkage findings of the BBF study are promising given that they either replicate confirmed linkage regions, e.g. 12q24.22-q24.32 or provide further support for linkage regions previously identified with suggestive evidence for linkage e.g. 2p23.1-p22.3. While the role of genetic variants driving the linkage signals in the BBF was not clarified in two case-control cohorts with BPD and unipolar depression, whole genome sequencing of the BBF is underway and will clarify the source of the linkage signals (if any), identify the genetic variants conferring susceptibility to BPD in the BBF, and elucidate their biological function. The hope for BPD and mood disorder genetic research is to one day translate clinical and genetic research into better diagnosis, prognosis and treatment with the ultimate aim being the alleviation of mood symptoms.

There is potential to bridge the gap between research and patients through drug development and gene therapy. Uncovering novel pathways involved in the aetiology of psychiatric diseases (e.g. for depression outside the monoaminergic pathway which is currently targeted by all known classes of antidepressants) (Skolnick, Popik, & Trullas, 2009) and better understanding already established pathways that could be targeted by new and possibly more effective therapeutic drugs is possible with the new wave of sequencing studies particularly those based on families segregating severe forms of psychiatric diseases and the collaborative atmosphere recently adopted in psychiatric genetics.

Appendix I

The spearman correlation matrix for the parametric linkage tests performed where 1=recessive, narrow, Branch 1, 2=recessive, narrow, BBF, 3=recessive, super, Branch 1, 4= recessive, super, BBF, 5=recessive, broad, Branch 1, 6=recessive, broad, BBF, 7= recessive, depression, Branch 1, 8=recessive, depression, BBF, 9=dominant, narrow, Branch 1, 10= dominant, narrow, BBF, 11= dominant, super, Branch 1, 12= dominant, super, BBF, 13= dominant, broad, Branch 1, 14= dominant, broad, BBF, 15= dominant, depression, Branch 1, 16= dominant, depression, BBF and the number of estimated independent tests using the Li and Ji (2005) method.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1.0000	0.9176	0.2868	0.2118	0.4505	0.4144	-0.2813	-0.2813	0.4635	0.3685	-0.5232	-0.5214	-0.3685	-0.3717	-0.3976	-0.4061
2	0.9176	1.0000	0.1505	0.1089	0.5551	0.5252	-0.3381	-0.3381	0.2966	0.2028	-0.2688	-0.2664	-0.4434	-0.4471	-0.4010	-0.4086
3	0.2868	0.1505	1.0000	0.9845	0.0652	0.0510	-0.0866	-0.0866	0.0567	0.0170	-0.1570	-0.1395	-0.1185	-0.1177	0.1216	0.1187
4	0.2118	0.1089	0.9845	1.0000	0.1152	0.1065	-0.0682	-0.0682	-0.0470	-0.0740	-0.1246	-0.1057	-0.0943	-0.0933	0.1999	0.1968
5	0.4505	0.5551	0.0652	0.1152	1.0000	0.9959	-0.1144	-0.1144	-0.5080	-0.5269	-0.1707	-0.1713	-0.1420	-0.1461	-0.1876	-0.1921
6	0.4144	0.5252	0.0510	0.1065	0.9959	1.0000	-0.1218	-0.1218	-0.5263	-0.5268	-0.1604	-0.1607	-0.1521	-0.1561	-0.1816	-0.1859
7	-0.2813	-0.3381	-0.0866	-0.0682	-0.1144	-0.1218	1.0000	1.0000	-0.2023	-0.2074	-0.0796	-0.0827	-0.0542	-0.0533	0.8288	0.8287
8	-0.2813	-0.3381	-0.0866	-0.0682	-0.1144	-0.1218	1.0000	1.0000	-0.2023	-0.2074	-0.0796	-0.0827	-0.0542	-0.0533	0.8288	0.8287
9	0.4635	0.2966	0.0567	-0.0470	-0.5080	-0.5263	-0.2023	-0.2023	1.0000	0.9747	-0.4418	-0.4424	-0.2749	-0.2739	-0.1844	-0.1871
10	0.3685	0.2028	0.0170	-0.0740	-0.5269	-0.5268	-0.2074	-0.2074	0.9747	1.0000	-0.4608	-0.4614	-0.2826	-0.2814	-0.1571	-0.1591
11	-0.5232	-0.2688	-0.1570	-0.1246	-0.1707	-0.1604	-0.0796	-0.0796	-0.4418	-0.4608	1.0000	0.9996	-0.1176	-0.1171	0.0234	0.0305
12	-0.5214	-0.2664	-0.1395	-0.1057	-0.1713	-0.1607	-0.0827	-0.0827	-0.4424	-0.4614	0.9996	1.0000	-0.1178	-0.1173	0.0277	0.0348
13	-0.3685	-0.4434	-0.1185	-0.0943	-0.1420	-0.1521	-0.0542	-0.0542	-0.2749	-0.2826	-0.1176	-0.1178	1.0000	0.9999	-0.1629	-0.1634
14	-0.3717	-0.4471	-0.1177	-0.0933	-0.1461	-0.1561	-0.0533	-0.0533	-0.2739	-0.2814	-0.1171	-0.1173	0.9999	1.0000	-0.1606	-0.1610
15	-0.3976	-0.4010	0.1216	0.1999	-0.1876	-0.1816	0.8288	0.8288	-0.1844	-0.1571	0.0234	0.0277	-0.1629	-0.1606	1.0000	0.9999
16	-0.4061	-0.4086	0.1187	0.1968	-0.1921	-0.1859	0.8287	0.8287	-0.1871	-0.1591	0.0305	0.0348	-0.1634	-0.1610	0.9999	1.0000

Original (total) number of variables (V) after removing redundant (collinear) variables:

16

USING THE REPORTEDLY MORE ACCURATE ESTIMATE OF THE Veff [VeffLi] PROPOSED BY LI AND JI (2005):

Effective Number of Independent Variables [VeffLi] (using Equation 5 of Li and Ji 2005):

8

Experiment-wide Significance Threshold Required to Keep Type I Error Rate at 5%:

0.00639115095454501

Appendix II

The spearman correlation matrix for the non-parametric linkage tests performed where 1=narrow, NPL_{pairs}, Branch 1, 2= super, NPL_{pairs}, Branch1, 3=broad, NPL_{pairs}, Branch 1, 4= depression, NPL_{pairs}, Branch 1, 5=narrow, NPL_{all}, Branch 1, 6=super, NPL_{all}, Branch1, 7= broad, NPL_{all}, Branch 1, 8=depression, NPL_{all}, Branch 1, 9=narrow, NPL_{pairs}, BBF, 10= super, NPL_{pairs}, BBF, 11=broad, NPL_{pairs}, BBF, 12= depression, NPL_{pairs}, BBF, 13=narrow, NPL_{all}, BBF, 14=super, NPL_{all}, BBF, 15= broad, NPL_{all}, BBF, 16=depression, NPL_{all}, BBF and the number of estimated independent tests using the Li and Ji (2005) method.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1.0000	-0.3720	0.8943	-0.7006	0.9708	-0.3334	0.9127	-0.5285	0.9415	-0.3633	0.8866	-0.5700	0.9665	-0.3293	0.9120	-0.4400
2	-0.3720	1.0000	-0.4739	0.2597	-0.3509	0.9594	-0.4942	-0.0828	-0.3218	0.9569	-0.4980	-0.0091	-0.3312	0.9515	-0.5284	-0.2025
3	0.8943	-0.4739	1.0000	-0.7239	0.8727	-0.4408	0.9826	-0.5320	0.7676	-0.4927	0.9589	-0.5949	0.8521	-0.4566	0.9836	-0.4494
4	-0.7006	0.2597	-0.7239	1.0000	-0.6554	0.2109	-0.7224	0.9314	-0.6320	0.1793	-0.7671	0.9464	-0.6479	0.1399	-0.7505	0.8727
5	0.9708	-0.3509	0.8727	-0.6554	1.0000	-0.3466	0.9177	-0.4837	0.9072	-0.3516	0.8717	-0.5229	0.9951	-0.3529	0.9126	-0.3925
6	-0.3334	0.9594	-0.4408	0.2109	-0.3466	1.0000	-0.4596	-0.1255	-0.3242	0.8936	-0.5019	-0.0739	-0.3417	0.9713	-0.4999	-0.2531
7	0.9127	-0.4942	0.9826	-0.7224	0.9177	-0.4596	1.0000	-0.5217	0.7720	-0.5256	0.9354	-0.5931	0.8915	-0.4898	0.9960	-0.4370
8	-0.5285	-0.0828	-0.5320	0.9314	-0.4837	-0.1255	-0.5217	1.0000	-0.4697	-0.1503	-0.5664	0.9825	-0.4802	-0.1960	-0.5395	0.9879
9	0.9415	-0.3218	0.7676	-0.6320	0.9072	-0.3242	0.7720	-0.4697	1.0000	-0.2314	0.8442	-0.4629	0.9354	-0.2541	0.7876	-0.3643
10	-0.3633	0.9569	-0.4927	0.1793	-0.3516	0.8936	-0.5256	-0.1503	-0.2314	1.0000	-0.4364	-0.0406	-0.3089	0.9550	-0.5384	-0.2449
11	0.8866	-0.4980	0.9589	-0.7671	0.8717	-0.5019	0.9354	-0.5664	0.8442	-0.4364	1.0000	-0.5917	0.8751	-0.4537	0.9560	-0.4599
12	-0.5700	-0.0091	-0.5949	0.9464	-0.5229	-0.0739	-0.5931	0.9825	-0.4629	-0.0406	-0.5917	1.0000	-0.5050	-0.1130	-0.6051	0.9728
13	0.9665	-0.3312	0.8521	-0.6479	0.9951	-0.3417	0.8915	-0.4802	0.9354	-0.3089	0.8751	-0.5050	1.0000	-0.3289	0.8913	-0.3838
14	-0.3293	0.9515	-0.4566	0.1399	-0.3529	0.9713	-0.4898	-0.1960	-0.2541	0.9550	-0.4537	-0.1130	-0.3289	1.0000	-0.5131	-0.3066
15	0.9120	-0.5284	0.9836	-0.7505	0.9126	-0.4999	0.9960	-0.5395	0.7876	-0.5384	0.9560	-0.6051	0.8913	-0.5131	1.0000	-0.4473
16	-0.4400	-0.2025	-0.4494	0.8727	-0.3925	-0.2531	-0.4370	0.9879	-0.3643	-0.2449	-0.4599	0.9728	-0.3838	-0.3066	-0.4473	1.0000

```
Original (total) number of variables (V) after removing redundant (collinear) variables:

16



USING THE REPORTEDLY MORE ACCURATE ESTIMATE OF THE Veff [VeffLi] PROPOSED BY LI AND JI (2005):

Effective Number of Independent Variables [VeffLi] (using Equation 5 of Li and Ji 2005):

5

Experiment-wide Significance Threshold Required to Keep Type I Error Rate at 5%:

0.0102062183130115
```

Appendix III

Gerome,

When you do parametric linkage analysis, it is possible to estimate a parameter alpha that estimate the fraction of all families with evidence for linkage. This doesn't actually translate in any simple way into the fraction of families with positive and negative LOD scores (technically, there is a distribution of per family likelihoods expected under the null and under the alternative and - in you data - Merlin estimated that the distribution seen across your sample is compatible to what would be expected under your linked model).

The delta parameter from the Kong and Cox LOD scores is not really interpretable - but if you want to know more, you should read the original paper by Augie Kong and Nancy Cox (1997, AJHG). It is pretty readable.

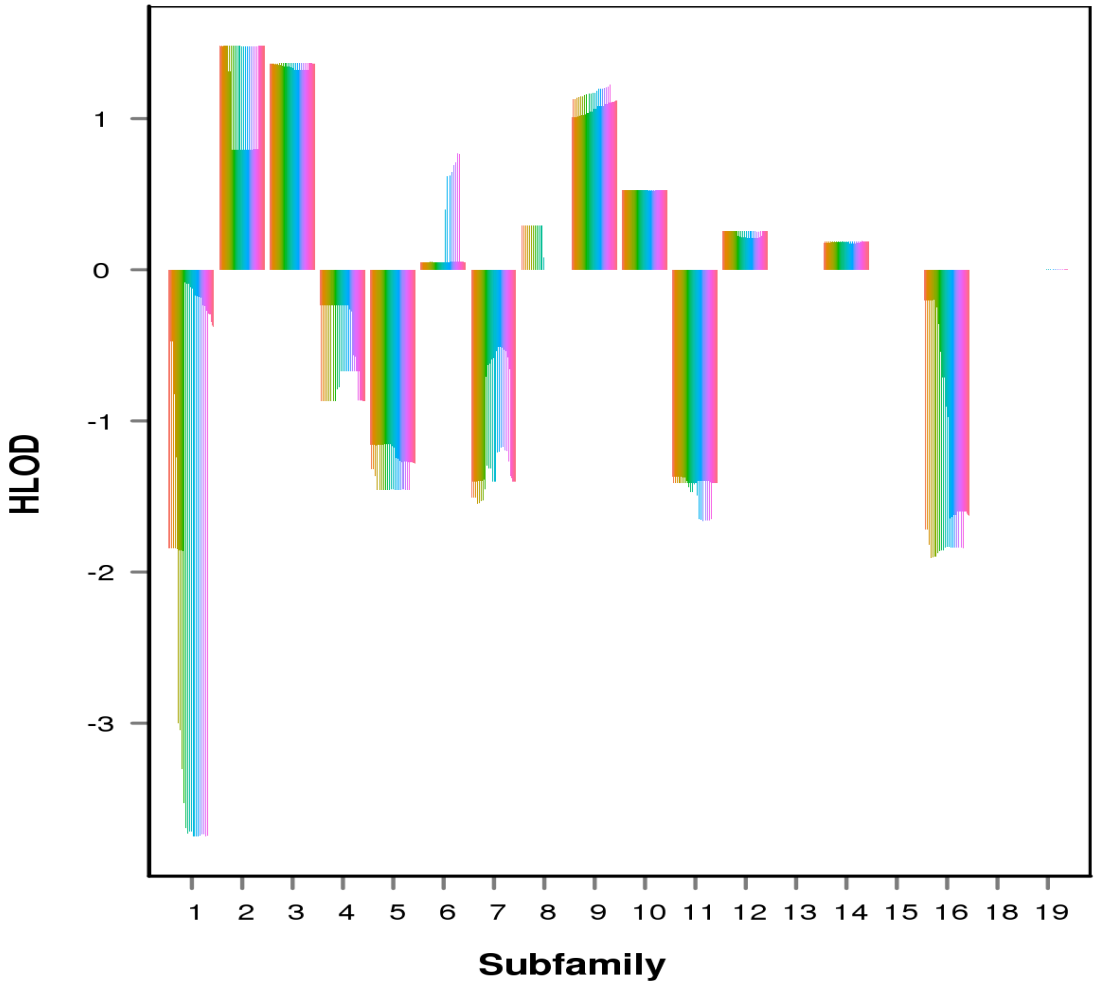
Gonçalo

Appendix IV

Full table for MERLIN near suggestive HLOD scores on maximum region (1-highest LOD of 2.47 region) on chromosome 3p25.1-p23.2 under the narrow phenotype model and the dominant mode of disease inheritance

SNP	Physical Position	cM Position	Branch 1		BBF	
			HLOD	α	HLOD	α
rs171239	10780817	28.48	2.47	0.49	2.48	0.49
rs2017892	10825748	28.55	2.47	0.49	2.48	0.49
rs2017903	10825837	28.55	2.47	0.49	2.48	0.49
rs720352	11056831	28.90	2.48	0.49	2.48	0.49
rs2616552	11666798	29.75	2.48	0.49	2.49	0.49
rs2030151	12749130	30.70	2.49	0.49	2.50	0.49
rs1828165	12895749	30.82	2.49	0.49	2.48	0.48
rs11080	13169322	31.27	2.49	0.50	2.45	0.54
rs59585	13501254	31.98	2.79	0.58	2.68	0.53
rs1368575	13831304	32.43	2.79	0.57	2.64	0.53
rs934448	13831352	32.43	2.79	0.57	2.64	0.50
rs3846122	14680270	33.75	2.65	0.55	2.49	0.49
rs1826215	14887182	34.11	2.60	0.54	2.44	0.47
rs1947147	15903002	35.84	2.60	0.54	2.40	0.47
rs723247	15997849	35.94	2.60	0.54	2.40	0.47
rs953245	16199191	36.16	2.59	0.54	2.39	0.47
rs1546377	16256377	36.24	2.58	0.54	2.38	0.45
rs723813	17240402	37.70	2.49	0.51	2.31	0.45
rs728022	17305303	37.74	2.48	0.50	2.30	0.43
rs958542	18247966	38.46	2.43	0.48	2.26	0.43
rs717793	18602534	38.91	2.42	0.48	2.26	0.43
rs717939	18626742	38.94	2.42	0.48	2.26	0.43
rs1398922	18670160	39.00	2.42	0.48	2.26	0.43
rs1372555	19371580	39.89	2.41	0.48	2.25	0.42
rs949665	19613756	40.19	2.41	0.47	2.25	0.42
rs1506101	19761201	40.38	2.40	0.47	2.24	0.42
rs1392573	20695652	41.56	2.38	0.47	2.23	0.42
rs904827	20696379	41.56	2.38	0.47	2.23	0.40
rs725542	21213394	42.22	2.33	0.45	2.18	0.40
rs721623	21650537	42.77	2.30	0.44	2.16	0.39
rs2358693	22574169	43.582	2.12	0.43	1.98	0.39
rs1074612	22579568	43.586	2.12	0.43	1.98	0.49
rs778480	23498492	44.75	1.67	0.41	1.54	0.37
rs2045300	23871402	45.34	1.70	0.43	1.56	0.38
rs951015	24053792	45.63	1.72	0.43	1.58	0.38
rs1112195	24085166	45.68	1.72	0.43	1.58	0.38
rs1394764	24405050	46.18	1.74	0.44	1.59	0.39
rs2196427	24533278	46.39	1.74	0.44	1.59	0.39
rs720822	25232155	47.77	1.94	0.47	1.77	0.42
rs1348979	27027497	49.44	2.12	0.49	1.94	0.44
rs2037472	27070331	49.49	2.12	0.49	1.95	0.44
rs724244	27242153	49.69	2.14	0.49	1.96	0.44
rs2370990	27690531	50.22	2.18	0.49	2.01	0.44
rs2371121	27850725	50.40	2.20	0.49	2.02	0.44
rs1609729	28303111	51.13	2.25	0.49	2.07	0.44
rs1979113	29057186	52.44	2.25	0.49	2.08	0.44

Appendix V



Subfamily contribution to the HLOD scores on chromosome 3p25.1-p23.2 as reported by all SNPs identified with near suggestive HLOD scores in MERLIN. SNPs in this region nearly achieved significance using MERLIN and were indicated with genome-wide significance by McLinkage. The Subfamilies 1 through to 12 constitute Branch 1, subfamilies 13 and 14 constitute Branch 2 and subfamilies 15 through to 19 constitute Branch 3. Subfamily 17 was dropped from the analyses.

Reference List

Abecasis, G. R., Cherny, S. S., & Cardon, L. R. (2001). The impact of genotyping error on family-based analysis of quantitative traits. *European Journal of Human Genetics*, *9*(2), 130-134.

Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature genetics*, *30*(1), 97-101.

Abecasis, G. R., & Wigginton, J. E. (2005). Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *The American Journal of Human Genetics*, *77*(5), 754-767.

Abkevich, V., Camp, N. J., Gutin, A., Farnham, J. M., Cannon-Albright, L., & Thomas, A. (2001). A robust multipoint linkage statistic (tlod) for mapping complex trait loci. *Genetic epidemiology*, *21*, S492-497.

Abkevich, V., Camp, N. J., Gutin, A., Farnham, J. M., Cannon-Albright, L., & Thomas, A. (2001). A robust multipoint linkage statistic (tlod) for mapping complex trait loci. *Genetic epidemiology*, *21*, S492.

Abkevich, V., Camp, N. J., Hensel, C. H., Neff, C. D., Russell, D. L., Hughes, D. C. et al. (2003). Predisposition locus for major depression at chromosome 12q22-12q23. 2. *The American Journal of Human Genetics*, *73*(6), 1271-1281.

Abou Jamra, R., Fuerst, R., Kaneva, R., Orozco Diaz, G., Rivas, F., Mayoral, F. et al. (2007). The first genomewide interaction and locus-heterogeneity linkage scan in bipolar affective disorder: strong evidence of epistatic effects between loci on chromosomes 2q and 6q. *The American Journal of Human Genetics*, *81*(5), 974-986.

Abreu, P. C., Greenberg, D. A., & Hodge, S. E. (1999). Direct power comparisons between simple LOD scores and NPL scores for linkage analysis in complex diseases. *The American Journal of Human Genetics*, *65*(3), 847-857.

Abreu, P. C., Hodge, S. E., & Greenberg, D. A. (2002). Quantification of type I error probabilities for heterogeneity LOD scores. *Genetic Epidemiology*, *22*(2), 156-169.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature protocols*, *5*(9), 1564-1573.

Akiskal, H. S. (2007). The emergence of the bipolar spectrum: validation along clinical-epidemiologic and familial-genetic lines. *Psychopharmacol Bull*, *40*(4), 99-115.

Akiskal, H. S., & Akiskal, K. (1992). Cyclothymic, hyperthymic and depressive temperaments as subaffective variants of mood disorders. *Annual Review*, *11*, 43-62.

American Psychiatric Association, A. P. (1994). *Diagnostic and Statistical Manual of Mental Disorders.* Washington DC: American Psychiatric Press.

Angst, J., Gamma, A., Benazzi, F., Ajdacic, V., Eich, D., & Rössler, W. (2003). Toward a re-definition of subthreshold bipolarity: epidemiology and proposed criteria for bipolar-II, minor bipolar disorders and hypomania. *Journal of Affective Disorders*, *73*(1-2), 133-146.

Angst, J., & Marneros, A. (2001). Bipolarity from ancient to modern times::: conception, birth and rebirth. *Journal of Affective Disorders*, *67*(1-3), 3-19.

Ardlie, K. G., Kruglyak, L., & Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, *3*(4), 299-309.

Arnold, L. M. (2003). Gender differences in bipolar disorder. *The Psychiatric Clinics of North America*, *26*(3), 595-620

Asherson, P., Walsh, C., Williams, J., Sargeant, M., Taylor, C., Clements, A. et al. (1994). Imprinting and anticipation. Are they relevant to genetic studies of schizophrenia? *The British Journal of Psychiatry*, *164*(5), 619-624.

Ayuso-Mateos, J. L., Vázquez-Barquero, J. L., Dowrick, C., Lehtinen, V., Dalgard, O. S., Casey, P. et al. (2001). Depressive disorders in Europe: prevalence figures from the ODIN study. *The British Journal of Psychiatry*, *179*(4), 308-316.

Badner, J. A., Gershon, E. S., & Goldin, L. R. (1998). Optimal ascertainment strategies to detect linkage to common disease alleles. *The American Journal of Human Genetics*, *63*(3), 880-888.

Badner, J. A., & Gershon, E. S. (2002). Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. *Molecular Psychiatry*, *7*(4), 405-411.

Badner, J. A., Koller, D., Foroud, T., Edenberg, H., Nurnberger, J. I., Zandi, P. P. et al. (2011). Genome-wide linkage analysis of 972 bipolar pedigrees using single-nucleotide polymorphisms. *Molecular Psychiatry*. E publication ahead of print.

Barden, N., Harvey, M., Gagné, B., Shink, E., Tremblay, M., Raymond, C. et al. (2006). Analysis of single nucleotide polymorphisms in genes in the chromosome 12Q24. 31 region points to P2RX7 as a susceptibility gene to bipolar affective disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *141*(4), 374-382.

Barnes, M. R. (2003). Genetic studies from genomic sequence. In M. R. Barnes, I. C. Gray, & J. Wiley (Eds.), *Bioinformatics for Geneticists* (pp. 179-202).

Baron, M. (2002). Manic-depression genes and the new millennium: poised for discovery. *Molecular Psychiatry*, *7*(4), 342-358.

Barrett, T. B., Hauger, R. L., Kennedy, J. L., Sadovnick, A. D., Remick, R. A., Keck, P. E. et al. (2003). Evidence that a single nucleotide polymorphism in the promoter of the G protein receptor kinase 3 gene is associated with bipolar disorder. *Molecular Psychiatry*, *8*(5), 546-557.

Beck, A. T., & Steer, R. A. (1984). Internal consistencies of the original and revised Beck Depression Inventory. *Journal of Clinical Psychology*, *40*(6), 1365-1367.

Benazzi, F. (1999). Prevalence and clinical features of atypical depression in depressed outpatients: a 467-case study. *Psychiatry Research*, *86*(3), 259-265.

Blackwelder, W. C., & Elston, R. C. (1982). Power and robustness of sib-pair linkage tests and extension to larger sibships. *Communications in Statistics-Theory and Methods*, *11*(5), 449-484.

Blackwood, D. H. R., Fordyce, A., Walker, M. T., St Clair, D. M., Porteous, D. J., & Muir, W. J. (2001). Schizophrenia and affective disorders—cosegregation with a translocation at chromosome 1q42 that directly disrupts brain-expressed genes: clinical and P300 findings in a family. *The American Journal of Human Genetics*, *69*(2), 428-433.

Blackwood, D., & Muir, W. (2001). Molecular genetics and the epidemiology of bipolar disorder. *Annals of Medicine*, *33*(4), 242-247.

Badano, J. L., & Katsanis, N. (2002). Beyond Mendel: an evolving view of human genetic disease transmission. *Nature Reviews Genetics*, *3*(10), 779-789.

Badner, J. A., & Gershon, E. S. (2002). Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. *Molecular psychiatry*, *7*(4), 405-411.

Baldessarini, R. J. (2000). A plea for integrity of the bipolar disorder concept. *Bipolar Disorders*, *2*(1), 3-7.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, *7*(10), 781-791.

Baron, M. (2002). Manic-depression genes and the new millennium: poised for discovery. *Molecular psychiatry*, *7*(4), 342-358.

Bellivier, F., Golmard, J. L., Rietschel, M., Schulze, T. G., Malafosse, A., Preisig, M. et al. (2003). Age at onset in bipolar I affective disorder: further evidence for three subgroups. *American Journal of Psychiatry*, *160*(5), 999-1001.

Bellivier, F., Golmard, J. L., Rietschel, M., Schulze, T. G., Malafosse, A., Preisig, M. et al. (2003). Age at onset in bipolar I affective disorder: further evidence for three subgroups. *American Journal of Psychiatry*, *160*(5), 999-1001.

Bodmer, W., & Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature genetics*, *40*(6), 695-701.

Bowden, D. W., An, S. S., Palmer, N. D., Brown, W. M., Norris, J. M., Haffner, S. M. et al. (2010). Molecular basis of a linkage peak: exome sequencing and family-based analysis identify a rare genetic variant in the ADIPOQ gene in the IRAS Family Study. *Human molecular genetics*, *19*(20), 4112-4120.

Bradwejn, J., Koszycki, D., du Tertre, A. C., van Megen, H., den Boer, J., Westenberg, H. et al. (1994). The panicogenic effects of cholecystokinin-tetrapeptide are antagonized by L-365,260, a central cholecystokinin receptor antagonist, in patients with panic disorder. *Archives of General Psychiatry*, *51*(6), 486-493.

Brasil, H., & Bordin, I. (2010). Convergent validity of K-SADS-PL by comparison with CBCL in a Portuguese speaking outpatient population. *BMC Psychiatry*, *10*(1), 83-94.

Breen, G., Webb, B. T., Butler, A. W., van den Oord, E. J. C. G., Tozzi, F., Craddock, N. et al. (2010). A genome-wide significant linkage for severe depression on chromosome 3: the Depression Network Study. *American Journal of Psychiatry*, *168*(8), 840-847.

Broman, K. W., & Weber, J. L. (2000). Characterization of human crossover interference. *The American Journal of Human Genetics*, *66*(6), 1911-1926.

Brzustowicz, L. M., Hodgkinson, K. A., Chow, E. W. C., Honer, W. G., & Bassett, A. S. (2000). Location of a major susceptibility locus for familial schizophrenia on chromosome 1q21-q22. *Science*, *288*(5466), 678-682.

Buetow, K. H. (1991). Influence of aberrant observations on high-resolution linkage analysis outcomes. *American journal of human genetics*, *49*(5), 985.

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A. et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661-678.

Buttenschön, H. N., Mors, O., Ewald, H., McQuillin, A., Kalsi, G., Lawrence, J. et al. (2004). No association between a neuronal nitric oxide synthase (NOS1) gene polymorphism on chromosome 12q24 and bipolar disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *124*(1), 73-75.

Cardno, A. G., Rijsdijk, F. V., Sham, P. C., Murray, R. M., & McGuffin, P. (2002). A twin study of genetic relationships between psychotic symptoms. *The American journal of psychiatry*, 159(4), 539-545.

Cardno, A. G., Rijsdijk, F. V., & West, R. M., Gottesmann, I.I, Craddock, N., Murray, R.M, & McGuffing, P. (2012). A twin study of schizoaffective‐mania, schizoaffective‐depression, and other psychotic syndromes. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetic.* E publication ahead of print.

Carr, I. M., Sheridan, E., Hayward, B. E., Markham, A. F., & Bonthron, D. T. (2009). IBDfinder and SNPsetter: Tools for pedigree‐independent identification of autozygous regions in individuals with recessive inherited disease. *Human Mutation*, *30*(6), 960-967.

Caspi, A., Sugden, K., Moffitt, T. E., Taylor, A., Craig, I. W., Harrington, H. L. et al. (2003). Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science*, *301*(5631), 386-389.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioural Research*, *1*(2), 245-276.

Cavalli-Sforza, L. L. (2001). *Genes, peoples, and languages.* Univ of California Pr.

Chang, K. (2009). Challenges in the diagnosis and treatment of pediatric bipolar depression. *Dialogues in Clinical Neuroscience*, *11*(1), 73-80.

Cheng, R., Juo, S. H., Loth, J. E., Nee, J., Iossifov, I., Blumenthal, R. et al. (2006). Genome-wide linkage scan in a large bipolar disorder sample from the National Institute of Mental Health genetics initiative suggests putative loci for bipolar disorder, psychosis, suicide, and panic disorder. *Molecular Psychiatry*, *11*(3), 252-260.

Chubb, J. E., Bradshaw, N. J., Soares, D. C., Porteous, D. J., & Millar, J. K. (2007). The DISC locus in psychiatric illness. *Molecular Psychiatry*, *13*(1), 36-64.

Cichon, S., Schumacher, J., Müller, D. J., Hürter, M., Windemuth, C., Strauch, K. et al. (2001). A genome screen for genes predisposing to bipolar affective disorder detects a new susceptibility locus on 8q. *Human Molecular Genetics*, *10*(25), 2933-2944

Cirulli, E. T., & Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, *11*(6), 415-425.

Clerget-Darpoux, F., Bonaïti-Pellié, C., & Hochez, J. (1986). Effects of misspecifying genetic parameters in lod score analysis. *Biometrics,* 393-399.

Cox, N. J., Frigge, M., Nicolae, D. L., Concannon, P., Hanis, C. L., Bell, G. I. et al. (1999). Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nature genetics*, *21*, 213-215.

Cordell, H. J. (2004). Bias toward the null hypothesis in model-free linkage analysis is highly dependent on the test statistic used. *The American Journal of Human Genetics*, *74*(6), 1294-1302.

Cordell, H. J. (2004). Bias toward the null hypothesis in model-free linkage analysis is highly dependent on the test statistic used. *The American Journal of Human Genetics*, *74*(6), 1294-1302.

Craddock, N., Owen, M., Burge, S., Kurian, B., Thomas, P., & McGuffin, P. (1994). Familial cosegregation of major affective disorder and Darier's disease (keratosis follicularis). *The British Journal of Psychiatry*, *164*(3), 355-358.

Craddock, N., Khodel, V., Van Eerdewegh, P., & Reich, T. (1995). Mathematical limits of multilocus models: the genetic transmission of bipolar disorder. *American Journal of Human Genetics*, *57*(3), 690-702.

Craddock, N., O'donovan, M. C., & Owen, M. J. (2005). The genetics of schizophrenia and bipolar disorder: dissecting psychosis. *Journal of Medical Genetics*, *42*(3), 193-204.

Craddock, N., & Forty, L. (2006). Genetics of affective (mood) disorders. *European Journal of Human Genetics*, *14*(6), 660-668.

Curtis, D., Kalsi, G., Brynjolfsson, J., McInnis, M., O'Neill, J., Smyth, C. et al. (2003). Genome scan of pedigrees multiply affected with bipolar disorder provides further support for the presence of a susceptibility locus on chromosome 12q23-q24, and suggests the presence of additional loci on 1p and 1q. *Psychiatric genetics*, *13*(2), 77-84.

Daly, M. J., Kruglyak, L., Pratt, S., Houstis, N., Reeve Mp, K. A., & Laner, E. S. (1998). GENEHUNTER 2.0-a complete linkage analysis system. *Am J Hum Genet Suppl*, *63*, A286.

Dawn Teare, M., & Barrett, J. H. (2005). Genetic linkage studies. *The Lancet*, *366*(9490), 1036-1044.

Dawson, E., Parfitt, E., Roberts, Q., Daniels, J., Lim, L., Sham, P. et al. (1995). Linkage studies of bipolar disorder in the region of the Darier's disease gene on chromosome 12q23‐24.1. *American journal of medical genetics*, *60*(2), 94-102.

Del Ben, C. M., Rodrigues, C. R., & Zuardi, A. W. (1996). Reliability of the Portuguese version of the structured clinical interview for DSM-III-R (SCID) in a Brazilian sample of psychiatric outpatients. *Brazilian Journal of Medical Biological Research*, *29*(12), 1675-1682.

Del-Favero, J., Gestel, S. V., Boerglum, A. D., Muir, W., Ewald, H., Mors, O. et al. (2002). European combined analysis of the CTG18. 1 and the ERDA1 CAG/CTG repeats in bipolar disorder. *European Journal of Human Genetics*, *10*(4), 276-280.

Del Zompo, M., Severino, G., Ardau, R., Chillotti, C., Piccardi, M., Dib, C. et al. (2010). Genome‐scan for bipolar disorder with sib‐pair families in the Sardinian population: A new susceptibility locus on chromosome 1p22–p21? *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *153*(6), 1200-1208.

Detera-Wadleigh, S. D., Badner, J. A., Berrettini, W. H., Yoshikawa, T., Goldin, L. R., Turner, G. et al. (1999). A high-density genome scan detects evidence for a bipolar-disorder susceptibility locus on 13q32 and other potential loci on 1q32 and 18p11. 2. *Proceedings of the National Academy of Sciences*, *96*(10), 5604-5609.

Dick, D. M., Foroud, T., Flury, L., Bowman, E. S., Miller, M. J., Rau, N. L. et al. (2003). Genomewide linkage analyses of bipolar disorder: a new sample of 250 pedigrees from the National Institute of Mental Health Genetics Initiative. *The American Journal of Human Genetics*, *73*(1), 107-114.

Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H., & Goldstein, D. B. (2010). Rare variants create synthetic genome-wide associations. *PLoS biology*, *8*(1).

Douglas, J. A., Skol, A. D., & Boehnke, M. (2002). Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *The American Journal of Human Genetics*, *70*(2), 487-495.

Dubicka, B., Carlson, G. A., Vail, A., & Harrington, R. (2008). Prepubertal mania: diagnostic differences between US and UK clinicians. *European Child & Adolescent Psychiatry*, *17*(3), 153-161.

Durner, M., & Greenberg, D. A. (1992). Effect of heterogeneity and assumed mode of inheritance on lod scores. *American Journal of Medical Genetics*, *42*(3), 271-275.

Dunteman, G. H. (1989). *Principal components analysis* (69). SAGE publications, Inc.

Edenberg, H. J., Foroud, T., Conneally, P. M., Sorbel, J. J., Carr, K., Crose, C. et al. (1997). Initial genomic scan of the NIMH genetics initiative bipolar pedigrees: chromosomes 3, 5, 15, 16, 17, and 22. *American Journal of Medical Genetics*, *74*(3), 238-246.

Ennis, S., Maniatis, N., & Collins, A. (2001). Allelic association and disease mapping. *Briefings in bioinformatics*, *2*(4), 375-387.

Ewald, H., Degn, B., Mors, O., & Kruse, T. A. (1998). Significant linkage between bipolar affective disorder and chromosome 12q24. *Psychiatric Genetics*, *8*(3):131-140.

Ewald, H., Flint, T., Kruse, T. A., & Mors, O. (2002). A genome-wide scan shows significant linkage between bipolar disorder and chromosome 12q24. 3 and suggestive linkage to chromosomes 1p22-21, 4p16, 6q14-22, 10q26 and 16p13. 3. *Molecular Psychiatry*, *7*(7), 734-744.

Ewald, H., Kruse, T. A., & Mors, O. (2003). Genome wide scan using homozygosity mapping and linkage analyses of a single pedigree with affective disorder suggests oligogenic inheritance. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *120*(1), 63-71.

Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, *29*(1), 51-76.

Fallin, M. D., Lasseter, V. K., Wolyniec, P. S., McGrath, J. A., Nestadt, G., Valle, D. et al. (2004). Genomewide linkage scan for bipolar-disorder susceptibility loci among Ashkenazi Jewish families. *The American Journal of Human Genetics*, *75*(2), 204-219.

Fallin, M. D., Lasseter, V. K., Avramopoulos, D., Nicodemus, K. K., Wolyniec, P. S., McGrath, J. A. et al. (2005). Bipolar I disorder and schizophrenia: a 440-single-nucleotide polymorphism screen of 64 candidate genes among Ashkenazi Jewish case-parent trios. *The American Journal of Human Genetics*, *77*(6), 918-936.

Faraone, S. V., Tsuang, M. T., & Tsuang, D. W. (1999). *The Basics: Epidemiologic Foundations of Psychiatric Genetics* (15-45). London: The Guilford Press.

Faraone, S. V., Lasky‐Su, J., Glatt, S. J., Eerdewegh, P. V., & Tsuang, M. T. (2006). Early onset bipolar disorder: possible linkage to chromosome 9q34. *Bipolar Disorders*, *8*(2), 144-151.

Farmer, A., Breen, G., Brewster, S., Craddock, N., Gill, M., Korszun, A. et al. (2004). The Depression Network (DeNT) Study: methodology and sociodemographic characteristics of the first 470 affected sibling pairs from a large multi-site linkage genetic study. *BMC psychiatry*, *4*(1), 42.

Farmer, A., Elkin, A., & McGuffin, P. (2007). The genetics of bipolar affective disorder. *Current Opinion in Psychiatry*, *20*(1), 8-12.

Farmer, A., Harris, T., Redman, K., Sadler, S., Mahmood, A., & McGuffin, P. (2000). Cardiff Depression Study. *The British Journal of Psychiatry*, *176*(2), 150-155.

Farrall, M. (1997). LOD wars: the affected-sib-pair paradigm strikes back! *American journal of human genetics*, *60*(3), 735-738.

Fatemi, S. H., Reutiman, T. J., Folsom, T. D., Bell, C., Nos, L., Fried, P. et al. (2006). Chronic olanzapine treatment causes differential expression of genes in frontal cortex of rats as revealed by DNA microarray technique. *Neuropsychopharmacology*, *31*(9), 1888-1899.

Ferreira, M. A. R., O'Donovan, M. C., Meng, Y. A., Jones, I. R., Ruderfer, D. M., Jones, L. et al. (2008). Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nature Genetics*, *40*(9), 1056-1058.

Fisfalen, M. E., Schulze, T. G., DePaulo, J. R., DeGroot, L. J., Badner, J. A., & McMahon, F. J. (2005). Familial variation in episode frequency in bipolar affective disorder. *American Journal of Psychiatry*, *162*(7), 1266-1272.

Frangos, E., Athanassenas, G., Tsitourides, S., Katsanou, N., & Alexandrakou, P. (1985). Prevalence of DSM III schizophrenia among the first‐degree relatives of schizophrenic probands. *Acta Psychiatrica Scandinavica*, *72*(4), 382-386.

Freimer, N. B., Sandkuijl, L. A., & Blower, S. M. (1993). Incorrect specification of marker allele frequencies: effects on linkage analysis. *American journal of human genetics*, *52*(6), 1102.

Fu, Y. H., Kuhl, D., Pizzuti, A., Pieretti, M., Sutcliffe, J. S., Richards, S. et al. (1991). Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell*, *67*(6), 1047-1058.

Fullerton, J. M., Tiwari, Y., Agahi, G., Heath, A., Berk, M., Mitchell, P. B. et al. (2010). Assessing oxidative pathway genes as risk factors for bipolar disorder. *Bipolar Disorders*, *12*(5), 550-556.

Fullerton, J. M., Donald, J. A., Mitchell, P. B., & Schofield, P. R. (2010). Two-dimensional genome scan identifies multiple genetic interactions in bipolar affective disorder. *Biological Psychiatry*, *67*(5), 478-486.

Gao, X., Becker, L. C., Becker, D. M., Starmer, J. D., & Province, M. A. (2010). Avoiding the high Bonferroni penalty in genome‐wide association studies. *Genetic Epidemiology*, *34*(1), 100-105.

Garner, C., McInnes, L. A., Spesny, M., Fournier, E., Leon, P., & Freimer, N. B. (2001). Linkage analysis of a complex pedigree with severe bipolar disorder, using a Markov chain Monte Carlo method. *The American Journal of Human Genetics*, *68*(4), 1061-1064.

Gaysina, D., Cohen, S., Craddock, N., Farmer, A., Hoda, F., Korszun, A. et al. (2008). No association with the 5, 10‐methylenetetrahydrofolate reductase gene and major depressive disorder: Results of the depression case control (DeCC) study and a meta‐analysis. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *147*(6), 699-706.

Gaysina, D., Cohen‐Woods, S., Chow, P. C., Martucci, L., Schosser, A., Ball, H. A. et al. (2009). Association of the dystrobrevin binding protein 1 gene (DTNBP1) in a bipolar case–control study (BACCS). *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *150*(6), 836-844.

Geller, B., Tillman, R., Bolhofner, K., Zimerman, B., Strauss, N. A., & Kaufmann, P. (2006). Controlled, blindly rated, direct-interview family study of a prepubertal and early-adolescent bipolar I disorder phenotype: morbid risk, age at onset, and comorbidity. *Archives of General Psychiatry*, *63*(10), 1130-1138.

Gershon, E. S., Hamovit, J., Guroff, J. J., Dibble, E., Leckman, J. F., Sceery, W. et al. (1982). A family study of schizoaffective, bipolar I, bipolar II, unipolar, and normal control probands. *Archives of General Psychiatry*, *39*(10), 1157.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H. et al. (2003). The international HapMap project. *Nature*, *426*(6968), 789-796.

Goldin, L. R., & Weeks, D. E. (1993). Two-locus models of disease: comparison of likelihood and nonparametric linkage methods. *American Journal of Human Genetics*, *53*(4), 908-915.

Göring, H. H. H., & Terwilliger, J. D. (2000). Linkage analysis in the presence of errors II: marker-locus genotyping errors modeled with hypercomplex recombination fractions. *The American Journal of Human Genetics*, *66*(3), 1107-1118.

Grant, B. F., Stinson, F. S., Hasin, D. S., Dawson, D. A., Chou, S. P., Ruan, W. et al. (2005). Prevalence, Correlates, and Comorbidity of Bipolar I Disorder and Axis I and II Disorders: Results From the National Epidemiologic Survey on Alcohol and Related Conditions. *Journal of Clinical Psychiatry*, 66(10), 1205-1215.

Green, E., Elvidge, G., Jacobsen, N., Glaser, B., Jones, I., O'Donovan, M. C. et al. (2005). Localization of bipolar susceptibility locus by molecular genetic analysis of the chromosome 12q23-q24 region in two pedigrees with bipolar disorder and Darier's disease. *American Journal of Psychiatry*, *162*(1), 35-42.

Green, E. K., Grozeva, D., Raybould, R., Elvidge, G., Macgregor, S., Craig, I. et al. (2009). P2RX7: A bipolar and unipolar disorder candidate susceptibility gene? *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *150*(8), 1063-1069.

Greenberg, D. A. (1989). Inferring mode of inheritance by comparison of lod scores. *American Journal of Medical Genetics*, *34*(4), 480-486.

Greenberg, D. A., Abreu, P., & Hodge, S. E. (1998). The power to detect linkage in complex disease by means of simple LOD-score analyses. *The American Journal of Human Genetics*, *63*(3), 870-879.

Griebel, G., Perrault, G., & Sanger, D. J. (1997). CCK receptor antagonists in animal models of anxiety: comparison between exploration tests, conflict procedures and a model based on defensive behaviours. *Behavioural Pharmacology*, *8*(6), 549-560.

Grigoroiu‐Serbanescu, M., Herms, S., Mühleisen, T. W., Georgi, A., Diaconu, C. C., Strohmaier, J. et al. (2009). Variation in P2RX7 candidate gene (rs2230912) is not associated with bipolar I disorder and unipolar major depression in four European samples. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *150*(7), 1017-1021.

Groves, J. O. (2007). Is it time to reassess the BDNF hypothesis of depression? *Molecular psychiatry*, *12*(12), 1079-1088.

Grozeva, D., Kirov, G., Ivanov, D., Jones, I. R., Jones, L., Green, E. K. et al. (2010). Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. *Archives of General Psychiatry*, *67*(4), 318-327.

Hasin, D. S., Stinson, F. S., Ogburn, E., & Grant, B. F. (2007). Prevalence, correlates, disability, and comorbidity of DSM-IV alcohol abuse and dependence in the United States: results from the National Epidemiologic Survey on Alcohol and Related Conditions. *Archives of General Psychiatry*, *64*(7), 830-842.

Hamshere, M. L., Bennett, P., Williams, N., Segurado, R., Cardno, A., Norton, N. et al. (2005). Genomewide linkage scan in schizoaffective disorder: significant evidence for linkage at 1q42 close to DISC1, and suggestive evidence at 22q11 and 19p13. *Archives of General Psychiatry*, *62*(10), 1081-1088.

Hardy, J. A., & Higgins, G. A. (1992). Alzheimer's disease: the amyloid cascade hypothesis. *Science, 256(5054), 184-185*.

Hashimoto, S., Ochs, R. L., Komiya, S., & Lotz, M. (1998). Linkage of chondrocyte apoptosis and cartilage degradation in human osteoarthritis. *Arthritis & Rheumatism*, *41*(9), 1632-1638.

Heun, R., & Maier, W. (1993). The distinction of bipolar II disorder from bipolar I and recurrent unipolar depression: results of a controlled family study. *Acta Psychiatrica Scandinavica*, *87*(4), 279-284.

Hejjas, K., Szekely, A., Domotor, E., Halmai, Z., Balogh, G., Schilling, B. et al. (2009). Association between depression and the Gln460Arg polymorphism of P2RX7 gene: a dimensional approach. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *150*(2), 295-299.

Hennah, W., Varilo, T., Kestilä, M., Paunio, T., Arajärvi, R., Haukka, J. et al. (2003). Haplotype transmission analysis provides evidence of association for DISC1 to schizophrenia and suggests sex-dependent effects. *Human Molecular Genetics*, *12*(23), 3151-3159.

Herzberg, I., Jasinska, A., Garcia, J., Jawaheer, D., Kremeyer, B., Duque, C. et al. (2006). Convergent linkage evidence from two Latin-American population isolates supports the presence of a susceptibility locus for bipolar disorder in 5q31–34. *Human molecular genetics*, *15*(21), 3146.

Hodge, S. E., Vieland, V. J., & Greenberg, D. A. (2002). HLODs remain powerful tools for detection of linkage in the presence of genetic heterogeneity. *American Journal of Human Genetics*, *70*(2), 556-559.

Hodgkinson, C. A., Goldman, D., Jaeger, J., Persaud, S., Kane, J. M., Lipsky, R. H. et al. (2004). Disrupted in schizophrenia 1 (DISC1): association with schizophrenia, schizoaffective disorder, and bipolar disorder. *The American Journal of Human Genetics*, *75*(5), 862-872.

Holden, A. L. (2002). The SNP consortium: summary of a private consortium effort to develop an applied map of the human genome. *Biotechniques*, *32*, S22-S26.

Hosang, G. M., Uher, R., Keers, R., Cohen-Woods, S., Craig, I., Korszun, A. et al. (2010). Stressful life events and the brain-derived neurotrophic factor gene in bipolar disorder. *Journal of Affective Disorders*, *125*(1-3), 345-349.

Huentelman, M., Craig, D., Shieh, A., Corneveaux, J., Hu-Lince, D., Pearson, J. et al. (2005). SNiPer: improved SNP genotype calling for Affymetrix 10K GeneChip microarray data. *BMC genomics*, *6*(1), 149.

Jackson, J. E., & Wiley, J. (1991). *A user's guide to principal components.* Wiley Online Library.

Jablonski, K. A., McAteer, J. B., de Bakker, P. I. W., Franks, P. W., Pollin, T. I., Hanson, R. L. et al. (2010). Common variants in 40 genes assessed for diabetes incidence and response to metformin and lifestyle intervention in the diabetes prevention program. *Diabetes*, *59*(10), 2672-2681.

Jasinska, A. J., Jawaheer, D., DeYoung, J., Levinson, M., Zhang, Z., Kremeyer, B. et al. (2009). A narrow and highly significant linkage signal for severe bipolar disorder in the chromosome 5q33 region in Latin American pedigrees. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *150*(7), 998-1006.

Johansen, C. T., Wang, J., Lanktree, M. B., Cao, H., McIntyre, A. D., Ban, M. R. et al. (2010). Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nature Genetics*, *42*(8), 684-687.

Jolliffe, I. T., & MyiLibrary. (2002). *Principal component analysis* (2). Wiley Online Library.

John, S., Shephard, N., Liu, G., Zeggini, E., Cao, M., Chen, W. et al. (2004). Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. *The American Journal of Human Genetics*, *75*(1), 54-64.

Jones, I., & Craddock, N. (2001). Candidate gene studies of bipolar disorder. *Annals of Medicine*, *33*(4), 248-256.

Jones, I., Jacobsen, N., Green, E. K., Elvidge, G. P., Owen, M. J., & Craddock, N. (2002). Evidence for familial cosegregation of major affective disorder and genetic markers flanking the gene for Darier's disease. *Molecular Psychiatry*, *7*(4), 424 427.

Jones, I., Kent, L., & Craddock, N. (2002). Genetics of affective disorders. In P. McGuffin, M. J. Owen, & I. I. Gottesman (Eds.), *Psychiatric genetics and genomics* (pp. 211-246). Oxford: Oxford University Press.

Judd, L. L., & Akiskal, H. S. (2003). The prevalence and disability of bipolar spectrum disorders in the US population: re-analysis of the ECA database taking into account subthreshold cases. *Journal of Affective Disorders*, *73*(1-2), 123-131.

Karg, K., Burmeister, M., Shedden, K., & Sen, S. (2011). The serotonin transporter promoter variant (5-HTTLPR), stress, and depression meta-analysis revisited: evidence of genetic moderation. *Archives of General Psychiatry*, *68*(5), 444-454.

Kassem, L., Lopez, V., Hedeker, D., Steele, J., Zandi, P., & McMahon, F. (2006). Familiality of polarity at illness onset in bipolar affective disorder. *American Journal of Psychiatry*, *163*(10), 1754-1759.

Kelsoe, J. R., Spence, M. A., Loetscher, E., Foguet, M., Sadovnick, A. D., Remick, R. A. et al. (2001). A genome survey indicates a possible susceptibility locus for bipolar disorder on chromosome 22. *Proceedings of the National Academy of Sciences*, *98*(2), 585-590.

Kelsoe, J. R. (2003). Arguments for the genetic basis of the bipolar spectrum. *Journal of affective disorders*, *73*(1-2), 183-197.

Kendler, K. S., Neale, M., Kessler, R., Heath, A., & Eaves, L. (1993a). A twin study of recent life events and difficulties. *Archives of General Psychiatry*, *50*(10), 789-796.

Kendler, K. S., McGuire, M., Gruenberg, A. M., O'Hare, A., Spellman, M., & Walsh, D. (1993b). The Roscommon Family Study: IV. Affective illness, anxiety

disorders, and alcoholism in relatives. *Archives of General Psychiatry*, *50*(12), 952.

Kendler, K. S., Pedersen, N. L., Neale, M. C., & Mathé, A. A. (1995). A pilot Swedish twin study of affective illness including hospital-and population-ascertained subsamples: results of model fitting. *Behavior Genetics*, *25*(3), 217-232.

Kendler, K. S. (1998). Major depression and the environment: A psychiatric genetic perspective. *Pharmacopsychiatry*, 31(1), 5-9.

Kendler, K. (2006). Reflections on the relationship between psychiatric genetics and psychiatric nosology. *American Journal of Psychiatry*, *163*(7), 1138-1146.

Kennedy, G. C., Matsuzaki, H., Dong, S., Liu, W., Huang, J., Liu, G. et al. (2003). Large-scale genotyping of complex DNA. *Nature biotechnology*, *21*(10), 1233-1237.

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S. et al. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: results from the National Comorbidity Survey. *Archives of General Psychiatry*, *51*(1), 8-19.

Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62*(6), 617-627.

Kieseppa, T., Partonen, T., Haukka, J., Kaprio, J., & Jouko, L. (2004). High concordance of bipolar I disorder in a nationwide sample of twins. *American Journal of Psychiatry*, *161*(10), 1814-1821.

Kirov, G., Grozeva, D., Norton, N., Ivanov, D., Mantripragada, K. K., Holmans, P. et al. (2009). Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. *Human Molecular Genetics*, *18*(8), 1497-1503

Klein, D. N., Depue, R. A., & Slater, J. F. (1985). Cyclothymia in the adolescent offspring of parents with bipolar affective disorder. *Journal of Abnormal Psychology*, *94*(2), 115-127.

Koefoed, P., Woldbye, D. P., Hansen, T. O., Hansen, E. S., Knudsen, G. M., Bolwig, T. G. et al. (2010). Gene variations in the cholecystokinin system in patients with panic disorder. *Psychiatric Genetics*, *20*(2), 59-64.

Kohn, Y., & Lerer, B. (2005). Excitement and confusion on chromosome 6q: the challenges of neuropsychiatric genetics in microcosm. *Molecular Psychiatry*, *10*(12), 1062-1073.

Kong, A., & Cox, N. J. (1997). Allele-sharing models: LOD scores and accurate linkage tests. *The American Journal of Human Genetics*, *61*(5), 1179-1188.

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G. et al. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics*, *40*(9), 1068-1075.

Kremeyer, B., García, J., Müller, H., Burley, M. W., Herzberg, I., Parra, M. V. et al. (2010). Genome-Wide Linkage Scan of Bipolar Disorder in a Colombian Population Isolate Replicates Loci on Chromosomes 7p21–22, 1p31, 16p12 and 21q21–22 and Identifies a Novel Locus on Chromosome 12q. *Human Heredity*, *70(4)*, 255-268.

Kruglyak, L., & Lander, E. S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics*, *57*(2), 439-454.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., & Lander, E. S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*, *58(6),* 1347-1636.

Kruglyak, L. (1997). The use of a genetic map of biallelic markers in linkage studies. *Nature Genetics*, *17*(1), 21-24.

Kunugi, H., Hashimoto, R., Okada, T., Hori, H., Nakabayashi, T., Baba, A. et al. (2006). Possible association between nonsynonymous polymorphisms of the anaplastic lymphoma kinase (ALK) gene and schizophrenia in a Japanese population. *Journal of Neural Transmission*, *113*(10), 1569-1573.

Kwon, E., & Wang, W. & Tsai, L.H. (2011). Validation of schizophrenia-associated genes CSMD1, C10orf26, CACNA1C and TCF4 as miR-137 targets. *Molecular Psychiatry*, E publication ahead of print.

Lambert, D., Middle, F., Hamshere, M. L., Segurado, R., Raybould, R., Corvin, A. et al. (2005). Stage 2 of the Wellcome Trust UK–Irish bipolar affective disorder sibling-pair genome screen: evidence for linkage on chromosomes 6q16–q21, 4q12–q21, 9p21, 10p14–p12 and 18q22. *Molecular Psychiatry*, *10*(9), 831-841.

Lander, E. S., & Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences*, *84*(8), 2363-2367.

Lander, E. S., & Kruglyak, K. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature genetics*, *11*, 241-247.

Lange, K., & Sobel, E. (1991). A random walk method for computing genetic location scores. *American journal of human genetics*, *49*(6), 1320.

Lange, K., Cantor, R., Horvath, S., Perola, M., Sabatti, C., Sinsheimer, J. et al. (2001). Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. *Am J Hum Genet*, *69*(Suppl 1), A1886.

Leboyer, M., Henry, C., Paillere‐Martinot, M. L., & Bellivier, F. (2005). Age at onset in bipolar affective disorders: a review. *Bipolar Disorders*, *7*(2), 111-118.

Leibenluft, E., & Rich, B. A. (2008). Pediatric bipolar disorder. *FOCUS: The Journal of Lifelong Learning in Psychiatry*, *6*(3), 331-347.

Leonhard, K. (1959). *The Classification of the Endogenous Psychoses.* (5th edition). New York NY: Wiley & Sons.

Lewis, C. M. (2002). Genetic association studies: design, analysis and interpretation. *Briefings in Bioinformatics*, *3*(2), 146-153.

Lewis, C. M., Levinson, D. F., Wise, L. H., DeLisi, L. E., Straub, R. E., Hovatta, I. et al. (2003). Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. *The American Journal of Human Genetics*, *73*(1), 34-48.

Lewis, C. M., Ng, M. Y., Butler, A. W., Cohen-Woods, S., Uher, R., Pirlo, K. et al. (2010). Genome-wide association study of major recurrent depression in the UK population. *American Journal of Psychiatry*, *167*(8), 949-957. Levy-Lahad, E., Wasco, W., Poorkaj, P., Romano, D. M., Oshima, J., Pettingell, W. H. et al. (1995). Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science*, *269*(5226), 973.

Li, J., & Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, *95*(3), 221-227.

Lins, T. C., Vieira, R. G., Abreu, B. S., Grattapaglia, D., & Pereira, R. W. (2010). Genetic composition of Brazilian population samples based on a set of twenty‐eight ancestry informative SNPs. *American Journal of Human Biology*, *22*(2), 187-192.

Lindblad, K., Nylander, P. O., Zander, C., Yuan, Q. P., Ståhle, L., Engström, C. et al. (1998). Two commonly expanded CAG/CTG repeat loci: involvement in affective disorders? *Molecular Psychiatry*, *3*(5), 405-410.

Lines, C., Challenor, J., & Traub, M. (1995). Cholecystokinin and anxiety in normal volunteers: an investigation of the anxiogenic properties of pentagastrin and reversal by the cholecystokinin receptor subtype B antagonist L-365,260. *British Journal of Clinical Pharmacology*, *39*(3), 235-242.

Liu, W., Di, X., Yang, G., Matsuzaki, H., Huang, J., Mei, R. et al. (2003). Algorithms for large-scale genotyping microarrays. *Bioinformatics*, *19*(18), 2397-2403.

Lucae, S., Salyakina, D., Barden, N., Harvey, M., Gagné, B., Labbé, M. et al. (2006). P2RX7, a gene coding for a purinergic ligand-gated ion channel, is associated with major depressive disorder. *Human Molecular Genetics*, *15*(16), 2438-2445.

MacDonald, M. E., Ambrose, C. M., Duyao, M. P., Myers, R. H., Lin, C., Srinidhi, L. et al. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, *72*(6), 971-983.

MacQueen, G. M., Hajek, T., & Alda, M. (2005). The phenotypes of bipolar disorder: relevance for genetic investigations. *Molecular Psychiatry*, *10*(9), 811-826.

Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G. et al. (1992). Myotonic dystrophy mutation: an unstable CTG repeat in the 3'untranslated region of the gene. *Science*, *255*(5049), 1253-1255.

Maier, W., Lichtermann, D., Minges, J., Hallmayer, J., Heun, R., Benkert, O. et al. (1993). Continuity and discontinuity of affective disorders and schizophrenia: results of a controlled family study. *Archives of General Psychiatry*, *50*(11), 871-883.

Manly, B. F. J. (2005). *Multivariate statistical methods: a primer*. Chapman & Hall/CRC.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J. et al. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747-753.

Marcheco‐Teruel, B., Flint, T. J., Wikman, F. P., Torralbas, M., Gonzalez, L., Blanco, L. et al. (2006). A genome‐wide linkage search for bipolar disorder susceptibility loci in a large and complex pedigree from the eastern part of Cuba. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *141*(8), 833-843.

Marchini, J., Cardon, L. R., Phillips, M. S., & Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature genetics*, *36*(5), 512-517.

Mathers, C., Fat, D. M., & Boerma, J. T. (2008). *The global burden of disease: 2004 update.* World Health Organization.

Mathews, C. A., & Reus, V. I. (2001). Assortative mating in the affective disorders: A systematic review and meta-analysis. *Comprehensive psychiatry*, *42*(4), 257-262.

McGuffin, P., Katz, R., & Aldrich, J. (1986). Past and present state examination: the assessment of 'lifetime ever'psychopathology. *Psychological Medicine*, *16*(02), 461-465.

McGuffin, P., & Katz, R. (1989). The genetics of depression and manic-depressive disorder. *The British Journal of Psychiatry*, *155*(3), 294-304.

McGuffin, P., Katz, R., & Bebbington, P. (1988). The Camberwell Collaborative Depression Study. III. Depression and adversity in the relatives of depressed probands. *The British Journal of Psychiatry*, *152*(6), 775-782.

McGuffin, P. (1991). Genetic models of madness. In P. McGuffin & R. Murray (Eds.), *The New Genetics of Mental Illness* (pp. 27-57). Oxford: Butterworth-Heinemann Ltd.

McGuffin, P., & Sargeant, M. P. (1991). Genetic markers and affective disorders. In P. McGuffin & R. Murray (Eds.), *The New Genetics of Mental Illness* (pp. 165-181). Oxford: Butterworth-Heinemann.

McGuffin, P., Rijsdijk, F., Andrew, M., Sham, P., Katz, R., & Cardno, A. (2003). The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Archives of General Psychiatry*, *60*(5), 497-502.

McGuffin, P., Knight, J., Breen, G., Brewster, S., Boyd, P. R., Craddock, N. et al. (2005). Whole genome linkage scan of recurrent depressive disorder from the depression network study. *Human molecular genetics*, *14*(22), 3337.

McInnes, L. A., Escamilla, M. A., Reus, V. I., Leon, P., Silva, S., Rojas, E. et al. (1996). A complete genome screen for genes predisposing to severe bipolar disorder in two Costa Rican pedigrees. *Proceedings of the National Academy of Sciences*, *93*(23), 13060.

McInnis, M. G., McMahon, F. J., Chase, G. A., Simpson, S. G., Ross, C. A., & DePaulo Jr, J. R. (1993). Anticipation in bipolar affective disorder. *American Journal of Human Genetics*, *53*(2), 385-390.

Mcintyre, R., Konarski, J., Misener, V., & Kennedy, S. (2005). Bipolar disorder and diabetes mellitus: epidemiology, etiology, and treatment implications. *Annals of Clinical Psychiatry 17* (2),83-93.

McMahon, F. J., & Schulze, T. G. (2005). Bipolar disorder as a phenotype in genetic studies. *Clinical Neuroscience Research*, *5*(1), 37-43.

McPeek, M. S. (1999). Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genetic Epidemiology*, *16*(3), 225-249.

McQueen, M. B., Devlin, B., Faraone, S. V., Nimgaonkar, V. L., Sklar, P., Smoller, J. W. et al. (2005). Combined analysis from eleven linkage studies of bipolar disorder provides strong evidence of susceptibility loci on chromosomes 6q and 8q. *The American Journal of Human Genetics*, *77*(4), 582-595.

McQuillin, A., Bass, N. J., Choudhury, K., Puri, V., Kosmin, M., Lawrence, J. et al. (2009). Case–control studies show that a non-conservative amino-acid change from a glutamine to arginine in the P2RX7 purinergic receptor protein is associated with both bipolar-and unipolar-affective disorders. *Molecular psychiatry*, *14*(6), 614-620.

McQuillin, A., Bass, N., Anjorin, A., Lawrence, J., Kandaswamy, R., Lydall, G. et al. (2011). Analysis of genetic deletions and duplications in the University College London bipolar disorder case control sample. *European Journal of Human Genetics*, *19*(5), 588-592.

McQuillin, A., Bass, N. J., Choudhury, K., Puri, V., Kosmin, M., Lawrence, J. et al. (2009). Case–control studies show that a non-conservative amino-acid change from a glutamine to arginine in the P2RX7 purinergic receptor protein is associated with both bipolar-and unipolar-affective disorders. *Molecular psychiatry*, *14*(6), 614-620.

Mendlewicz, J., & Rainer, J. D. (1977). Adoption study supporting genetic transmission in manic-depressive illness. *Nature*, *268*(5618), 327-329.

Merikangas, K., & Yu, K. (2002). Genetic epidemiology of bipolar disorder. *Clinical Neuroscience Research*, *2*(3), 127-141.

Merikangas, K. R., Akiskal, H. S., Angst, J., Greenberg, P. E., Hirschfeld, R., Petukhova, M. et al. (2007). Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey replication. *Archives of General Psychiatry*, *64*(5), 543-552.

Middleton, F. A., Pato, M. T., Gentile, K. L., Morley, C. P., Zhao, X., Eisener, A. F. et al. (2004). Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide–polymorphism (SNP) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. *The American Journal of Human Genetics*, *74*(5), 886-897.

Millar, J. K., Wilson-Annan, J. C., Anderson, S., Christie, S., Taylor, M. S., Semple, C. A. M. et al. (2000). Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Human Molecular Genetics*, *9*(9), 1415-1423.

Morissette, J., Villeneuve, A., Bordeleau, L., Rochette, D., Laberge, C., Gagne, B. et al. (1999). Genome-wide search for linkage of bipolar affective disorders in a very large pedigree derived from a homogeneous population in Quebec points to a locus of major effect on chromosome 12q23-q24. *American Journal of Medical Genetics*, *88*(5), 567-587.

Morton, N. E. (1955). Sequential tests for the detection of linkage. *American Journal of Human Genetics*, *7*(3), 277.

Muglia, P., Tozzi, F., Galwey, N. W., Franks, C., Upmanyu, R., Kong, X. Q. et al. (2010). Genome-wide association sutdy of recurrent major depressive disorder in two European case-control cohorts. *Molecular Psychiatry*, *15*, 589-601.

Murphy, K. C., Jones, L. A., & Owen, M. J. (1999). High rates of schizophrenia in adults with velo-cardio-facial syndrome. *Archives of General Psychiatry*, *56*(10), 940.

Myers, S., Bottolo, L., Freeman, C., McVean, G., & Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, *310*(5746), 321.

North, B. V., Curtis, D., & Sham, P. C. (2002). A note on the calculation of empirical P values from Monte Carlo procedures. *American Journal of Human Genetics*, *71*(2), 439-441.

O'Donovan, M. C., & Owen, M. J. (1996). Dynamic mutations and psychiatric genetics. *Psychological Medicine*, *26*(01), 1-6.

O'Donovan, M., Jones, I., & Craddock, N. (2003). *Anticipation and repeat expansion in bipolar disorder*. Proceedings from American Journal of Medical Genetics Part C: Seminars in Medical Genetics.

O'Donovan, M. C., Craddock, N. J., & Owen, M. J. (2009). Genetics of psychosis; insights from views across the genome. *Human genetics*, *126*(1), 3-12.

Osby, U., Brandt, L., Correia, N., Ekbom, A., & Sparen, P. (2001). Excess mortality in bipolar and unipolar disorder in Sweden. *Archives of General Psychiatry*, *58*(9), 844.

Ott, J. (1999). *Analysis of human genetic linkage.* Johns Hopkins Univ Pr.

Penrose, L. S., & Watson, E. M. (1945). A sex-linked tendency in familial diabetes. *Proc Am Diabetes Assoc*, *5*, 165-177.

Ott, J., & Rabinowitz, D. (1997). The effect of marker heterozygosity on the power to detect linkage disequilibrium. *Genetics*, *147*(2), 927-930.

McQuillin, A., Bass, N. J., Choudhury, K., Puri, V., Kosmin, M., Lawrence, J. et al. (2009). Case–control studies show that a non-conservative amino-acid change from a glutamine to arginine in the P2RX7 purinergic receptor protein is associated with both bipolar-and unipolar-affective disorders. *Molecular psychiatry*, *14*(6), 614-620.

Paddock, S., Laje, G., Charney, D., Rush, A., Wilson, A., Sorant, A. et al. (2007). Association of GRIK4 with outcome of antidepressant treatment in the STAR* D cohort. *American Journal of Psychiatry*, *164*(8), 1181-1188.

Palmatier, M. A., Pakstis, A. J., Speed, W., Paschou, P., Goldman, D., Odunsi, A. et al. (2004). COMT haplotypes suggest P2 promoter region relevance for schizophrenia. *Molecular psychiatry*, *9*(9), 859-870.

Pato, C. N., Pato, M. T., Kirby, A., Petryshen, T. L., Medeiros, H., Carvalho, C. et al. (2004). Genome‐wide scan in Portuguese Island families implicates multiple loci in bipolar disorder: Fine mapping adds support on chromosomes 6 and 11. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *127*(1), 30-34.

Payami, H., Kaye, J., Heston, L. L., Bird, T. D., & Schellenberg, G. D. (1993). Apolipoprotein E genotype and Alzheimer's disease. *Lancet*, *342*(8873), 738.

Pekkarinen, P., Terwilliger, J., Bredbacka, P. E., Lönnqvist, J., & Peltonen, L. (1995). Evidence of a predisposing locus to bipolar disorder on Xq24-q27. 1 in an extended Finnish pedigree. *Genome Research*, *5*(2), 105-115.

Pergadia, M. L., Glowinski, A. L., Wray, N. R., Agrawal, A., Saccone, S. F., Loukola, A. et al. (2011). A 3p26-3p25 genetic linkage finding for DSM-IV major depression in heavy smoking families. *American Journal of Psychiatry*, *168*(8), 848-852.

Pericak-Vance, M. A., Bebout, J. L., Gaskell, P. C., Yamaoka, L. H., Hung, W. Y., Alberts, M. J. et al. (1991). Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. *American Journal of Human Genetics*, *48*(6), 1034-1050.

Potash, J. B., Willour, V. L., Chiu, Y. F., Simpson, S. G., MacKinnon, D. F., Pearlson, G. D. et al. (2001). The familial aggregation of psychotic symptoms in bipolar disorder pedigrees. *American Journal of Psychiatry*, *158*(8), 1258-1264.

Post, R. M., Luckenbaugh, D. A., Leverich, G. S., Altshuler, L. L., Frye, M. A., Suppes, T. et al. (2008). Incidence of childhood-onset bipolar illness in the USA and Europe. *The British Journal of Psychiatry*, *192*(2), 150-151.

Prata, D. P., Mechelli, A., Picchioni, M., Fu, C. H. Y., Kane, F., Kalidindi, S. et al. (2010). No association of Disrupted-in-Schizophrenia-1 variation with prefrontal function in patients with schizophrenia and bipolar disorder. *Genes, Brain and Behaviour, 10*(3), 276-285.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, *38*(8), 904-909.

Prochazka, M., Farook, V. S., Ossowski, V., Wolford, J. K., & Bogardus, C. (2002). Variant screening of PRKAB2, a type 2 diabetes mellitus susceptibility candidate gene on 1q in Pima Indians. *Molecular and cellular probes*, *16*(6), 421-427.

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P. et al. (2010). LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, *26*(18), 2336-2337.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D. et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559-575.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D. et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559-575.

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F. et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), 748-752.

Reich, T., Cloninger, C. R., & Guze, S. B. (1975). The multifactorial model of disease transmission: I. Description of the model and its use in psychiatry. *The British Journal of Psychiatry*, *127*(1), 1-10.

Reich, T., Clayton, P. J., & Winokur, G. (1969). Family history studies: V. The genetics of mania. *The American Journal of psychiatry*.

Reich, D. E., & Goldstein, D. B. (2001). Detecting association in a case-control study while correcting for population stratification. *Genetic epidemiology*, *20*(1), 4-16.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J. et al. (2001). Linkage disequilibrium in the human genome. *Nature*, *411*(6834), 199-204.

Rice, J., Reich, T., Andreasen, N. C., Endicott, J., Van Eerdewegh, M., Fishman, R. et al. (1987). The familial transmission of bipolar illness. *Archives of General Psychiatry*, *44*(5), 441-447.

Rijsdijk, F. V., Sham, P. C., Sterne, A., Purcell, S., McGuffin, P., Farmer, A. et al. (2001). Life events and depression in a community sample of siblings. *Psychological Medicine*, *31*(03), 401-410.

Ripke, S., Sanders, A. R., Kendler, K. S., Levinson D. F. et al. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics, 43*(10), 969-876.

Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, *273*(5281), 1516-1517.

Risch, N., & Giuffra, L. (1992). Model misspecification and multipoint linkage analysis. *Human Heredity*, *42*(1), 77-92.

Roach, J. C., Glusman, G., Smit, A. F. A., Huff, C. D., Hubley, R., Shannon, P. T. et al. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, *328*(5978), 636-639.

Roberts, S. B., MacLean, C. J., Neale, M. C., Eaves, L. J., & Kendler, K. S. (1999). Replication of linkage studies of complex traits: an examination of variation in location estimates. *The American Journal of Human Genetics*, *65*(3), 876-884.

Rogaev, E. I., Sherrington, R., Rogaeva, E. A., Levesque, G., Ikeda, M., Liang, Y. et al. (1995). Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature*, *376*(6543), 775-778.

Romeo, S., Yin, W., Kozlitina, J., Pennacchio, L. A., Boerwinkle, E., Hobbs, H. H. et al. (2009). Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *The Journal of Clinical Investigation*, *119*(1), 70.

Rotondo, A., Mazzanti, C., Dellâ, L., Rucci, P., Sullivan, P., Bouanani, S. et al. (2002). Catechol o-methyltransferase, serotonin transporter, and tryptophan hydroxylase gene polymorphisms in bipolar disorder patients with and without comorbid panic disorder. *American Journal of Psychiatry*, *159*(1), 23-29.

Rudan, I., Rudan, D., Campbell, H., Carothers, A., Wright, A., Smolej-Narancic, N. et al. (2003). Inbreeding and risk of late onset complex disease. *Journal of Medical Genetics*, *40*(12), 925-932.

Sakuntabhai, A., Ruiz-Perez, V., Carter, S., Jacobsen, N., Burge, S., Monk, S. et al. (1999). Mutations in ATP2A2, encoding a Ca2+ pump, cause Darier disease. *Nature Genetics*, *21*(3), 271-277.

Saunders, E. H., Scott, L. J., McInnis, M. G., & Burmeister, M. (2008). Familiality and diagnostic patterns of subphenotypes in the National Institutes of Mental Health bipolar sample. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *147*(1), 18-26.

Schalling, M., Hudson, T. J., Buetow, K. H., & Housman, D. E. (1993). Direct detection of novel expanded trinucleotide repeats in the human genome. *Nature Genetics*, *4*(2), 135-139.

Schmidt, S. (2006). Advanced parametric linkage analysis. In *Genetic Analysis of Complex Diseases* (pp. 255-281). Wiley Online Library.

Schork, N. J., & Greenwood, T. A. (2004). Inherent bias toward the null hypothesis in conventional multipoint nonparametric linkage analysis. *The American Journal of Human Genetics*, *74*(2), 306-316.

Schosser, A., Cohen‐Woods, S., Gaysina, D., Chow, P. C., Martucci, L., Farmer, A. et al. (2010). NRG1 gene in recurrent major depression: No association in a large‐scale case–control association study. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *153*(1), 141-147.

Schulze, T. G., Hedeker, D., Zandi, P., Rietschel, M., & McMahon, F. J. (2006). What is familial about familial bipolar disorder?: Resemblance among relatives across a broad spectrum of phenotypic characteristics. *Archives of general psychiatry*, *63*(12), 1368-1376.

Schumacher, J., Kaneva, R., Jamra, R. A., Diaz, G. O., Ohlraun, S., Milanova, V. et al. (2005). Genomewide scan and fine-mapping linkage studies in four European samples with bipolar affective disorder suggest a new susceptibility locus on chromosome 1p35-p36 and provides further evidence of loci on chromosome 4q31 and 6q24. *The American Journal of Human Genetics*, *77*(6), 1102-1111.

Schwab, S. G., Handoko, H. Y., Kusumawardhani, A., Widyawati, I., Amir, N., Nasrun, M. W. S. et al. (2008). Genome‐wide scan in 124 Indonesian sib‐pair families with schizophrenia reveals genome‐wide significant linkage to a locus on chromosome 3p26‐21. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *147*(7), 1245-1252.

Scott, A. L. B. W. K., Schmidt, E. R. M. S., Ashley-Koch, Y. J. L. A., & Schmidt, M. P. B. M. (2005). Linkage disequilibrium inflates type I error rates in multipoint linkage analysis when parental genotypes are missing. *Human Heredity*, *59*(4), 220-227.

Service, S., Molina, J., DeYoung, J., Jawaheer, D., Aldana, I., Vu, T., Araya, C., Araya X. et al. (2006). Results of a SNP genome screen in a large Costa Rican pedigree segregating for severe bipolar disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetic,141(4),367-373.*

Segurado, R., Detera-Wadleigh, S. D., Levinson, D. F., Lewis, C. M., Gill, M., Nurnberger, J. I. et al. (2003). Genome scan meta-analysis of schizophrenia and bipolar disorder, part III: Bipolar disorder. *The American Journal of Human Genetics*, *73*(1), 49-62.

Serretti, A., & Mandelli, L. (2008). The genetics of bipolar disorder: genome 'hot regions,'genes, new potential candidates and future directions. *Molecular psychiatry*, *13*(8), 742-771.

Sham, P. (1996). Genetic epidemiology. *British medical bulletin*, *52*(3), 408.

Sham, P., & McGuffin, P. (2002). Linkage and Association. In P. McGuffin, M. J. Owen, & I. I. Gottesman (Eds.), *Psychiatric genetics and genomics* (pp. 55-73). Oxford University Press, USA.

Sham, S., & Zhao, J. (1998). Linkage analysis using affected sib pairs. In M. Bishop (Ed.), *Guide to human genome computing* (pp. 74-87). Academic Press.

Sherrington, R., Rogaev, E. I., Liang, Y., Rogaeva, E. A., Levesque, G., Ikeda, M. et al. (1995). Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature*, *375*(6534), 754-760.

Shifman, S., Bronstein, M., Sternfeld, M., Pisanté-Shalom, A., Lev-Lehman, E., Weizman, A. et al. (2002). A highly significant association between a COMT haplotype and schizophrenia. *The American Journal of Human Genetics*, *71*(6), 1296-1302.

Shink, E., Harvey, M., Tremblay, M., Gagné, B., Belleau, P., Raymond, C. et al. (2005). Analysis of microsatellite markers and single nucleotide polymorphisms in candidate genes for susceptibility to bipolar affective disorder in the chromosome 12Q24. 31 region. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *135*(1), 50-58.

Shink, E., Morissette, J., & Barden, N. (2003). Genetic heterogeneity in a very large bipolar affective disorder pedigree from Quebec. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *119*(1), 65-68.

Shyn, S. I., Shi, J., Kraft, J. B., Potash, J. B., Knowles, J. A., Weissman, M. M. et al. (2011). Novel loci for major depression identified by genome-wide association study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of three studies. *Molecular Psychiatry*, *16*(2), 202-215.

Silberberg, G., Ben‐Shachar, D., & Navon, R. (2010). Genetic analysis of nitric oxide synthase 1 variants in schizophrenia and bipolar disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *153*(7), 1318-1328.

Sklar, P., Ripke, S., & Scott, L. J. Andreassen, O.A., Cichon, S. et al. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature , 43(10), 977-983.*

Sklar, P., Smoller, J. W., Fan, J., Ferreira, M. A. R., Perlis, R. H., Chambert, K. et al. (2008). Whole-genome association study of bipolar disorder. *Molecular Psychiatry*, *13*(6), 558-569.

Skolnick, P., Popik, P., & Trullas, R. (2009). Glutamate-based antidepressants: 20 years on. *Trends in Pharmacological Sciences*, *30*(11), 563-569.

Smith, D. J., Griffiths, E., Kelly, M., Hood, K., Craddock, N., & Simpson, S. A. (2011). Unrecognised bipolar disorder in primary care patients with depression. *British Journal of Psychiatry*, *199*(1), 49-56.

Smith, E. N., Bloss, C. S., Badner, J. A., Barrett, T., Belmonte, P. L., Berrettini, W. et al. (2009). Genome-wide association study of bipolar disorder in European American and African American individuals. *Molecular Psychiatry*, *14*(8), 755-763.

Smoller, J. W., & Finn, C. T. (2003). Family, twin, and adoption studies of bipolar disorder. *Proceedings from American Journal of Medical Genetics Part C: Seminars in Medical Genetics, 123*(1), 48-58.

Smrt, R. D., Szulwach, K. E., Pfeiffer, R. L., Li, X., Guo, W., Pathania, M. et al. (2010). MicroRNA miR‐137 Regulates Neuronal Maturation by Targeting Ubiquitin Ligase Mind Bomb‐1. *Stem Cells*, *28*(6), 1060-1070.

Soares, J. C., & Gershon, S. (2000). The diagnostic boundaries of bipolar disorder. *Bipolar Disorders*, *2*(1), 1-2.

Sobel, E., & Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics*, *58*(6), 1323-1337.

Sobel, E., Papp, J. C., & Lange, K. (2002). Detection and integration of genotyping errors in statistical genetics. *The American Journal of Human Genetics*, *70*(2), 496-508.

Spence, M. A., Flodman, P. L., Sadovnick, A. D., Bailey‐Wilson, J. E., Ameli, H., & Remick, R. A. (1995). Bipolar disorder: evidence for a major locus. *American journal of Medical Genetics*, *60*(5), 370-376.

St Clair, D., Blackwood, D., Muir, W., Walker, M., Carothers, A., Spowart, G. et al. (1990). Association within a family of a balanced autosomal translocation with major mental illness. *The Lancet*, *336*(8706), 13-16.

Stefansson, H., Ophoff, R. A., Steinberg, S., Andreassen, O. A., Cichon, S., Rujescu, D. et al. (2009). Common variants conferring risk of schizophrenia. *Nature*, *460*(7256), 744-747.

Stine, O. C., McMahon, F. J., Chen, L., Xu, J., Meyers, D. A., MacKinnon, D. F. et al. (1997). Initial Genome Screen for Bipolar Disorder in the NIMH Genetics Initiative Pedigrees: Chromosomes 2. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *74*(3), 263-269.

Suarez, B. K., Hampe, C. L., & Van Eerdewegh, P. (1994). Problems of replicating linkage claims in psychiatry. *Genetic Approaches to Mental Disorders,* 23-46.

Sullivan, P. F., de Geus, E. J. C., Willemsen, G., James, M. R., Smit, J. H., Zandbelt, T. et al. (2008). Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo. *Molecular Psychiatry*, *14*(4), 359-375.

Svenningsson, P., Chergui, K., Rachleff, I., Flajolet, M., Zhang, X., Yacoubi, M. E. et al. (2006). Alterations in 5-HT1B receptor function by p11 in depression-like states. *Science*, *311*(5757), 77-80.

Team, R. (2010). R: A language and environment for statistical computing. *R Foundation for Statistical Computing Vienna Austria*.

Terwilliger, J. D., & Ott, J. (1994). *Handbook of human genetic linkage.* Johns Hopkins Univ Pr.

Terwilliger, J. D., Shannon, W. D., Lathrop, G. M., Nolan, J. P., Goldin, L. R., Chase, G. A. et al. (1997). True and false positive peaks in genomewide scans: applications of length-biased sampling to linkage mapping. *The American Journal of Human Genetics*, *61*(2), 430-438.

Thomas, A., & Camp, N. J. (2006). Maximum likelihood estimates of allele frequencies and error rates from samples of related individuals by gene counting. *Bioinformatics*, *22*(6), 771-772.

Thomas, A., Camp, N. J., Farnham, J. M., Allen‐Brady, K., & Cannon‐Albright, L. A. (2008). Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays. *Annals of human genetics*, *72*(2), 279-287.

Tsuang, M. T., & Dempsey, G. M. (1979). Long-term outcome of major psychoses: II. Schizoaffective disorder compared with schizophrenia, affective disorders, and a surgical control group. *Archives of general psychiatry*, *36*(12), 1302.

Tsui, L. C. (1992). The spectrum of cystic fibrosis mutations. *Trends in Genetics*, *8*(11), 392-398.

Uher, R., & McGuffin, P. (2008). The moderation by the serotonin transporter gene of environmental adversity in the aetiology of mental illness: review and methodological analysis. *Molecular Psychiatry*, *13*(2), 131-146.

Uher, R., Huezo-Diaz, P., Perroud, N., Smith, R., Rietschel, M., Mors, O. et al. (2009). Genetic predictors of response to antidepressants in the GENDEP project. *The pharmacogenomics Journal*, *9*(4), 225-233.

Uher, R., Perroud, N., Ng, M. Y. M., Hauser, J., Henigsberg, N., Maier, W. et al. (2010). Genome-wide pharmacogenetics of antidepressant response in the GENDEP project. *American Journal of Psychiatry*, *167*(5), 555-564.

Venken, T., Claes, S., Sluijs, S., Paterson, A. D., Van Duijn, C., Adolfsson, R. et al. (2005). Genomewide scan for affective disorder susceptibility loci in families of a northern Swedish isolated population. *The American Journal of Human Genetics*, *76*(2), 237-248.

Vincent, J. B., Paterson, A. D., Strong, E., Petronis, A., & Kennedy, J. L. (2000). The unstable trinucleotide repeat story of major psychosis. *American Journal of Medical Genetics*, *97*(1), 77-97.

Visscher, P. M., Goddard, M. E., Derks, E. M., & Wray, N. R. (2011). Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. *Molecular Psychiatry*. E publication ahead of pring.

von Knorring, A. L., Cloninger, C. R., Bohman, M., & Sigvardsson, S. (1983). An adoption study of depressive disorders and substance abuse. *Archives of general psychiatry*, *40*(9), 943-950.

Weissman, M. M., Leaf, P. J., Tischler, G. L., Blazer, D. G., Karno, M., Bruce, M. L. et al. (1988). Affective disorders in five United States communities. *Psychological Medicine*, *18*(01), 141-153.

Weissman, M. M., Bland, R. C., Canino, G. J., Faravelli, C., Greenwald, S., Hwu, H. G. et al. (1996). Cross-national epidemiology of major depression and bipolar disorder. *JAMA: the Journal of the American Medical Association*, *276*(4), 293.

Wender, P. H., Kety, S. S., Rosenthal, D., Schulsinger, F., Ortmann, J., & Lunde, I. (1986). Psychiatric disorders in the biological and adoptive families of adopted individuals with affective disorders. *Archives of General Psychiatry*, *43*(10), 923.

Whittemore, A. S., & Halpern, J. (1994). A class of tests for linkage using affected pedigree members. *Biometrics,* 118-127.

Wigginton, J. E., & Abecasis, G. R. (2005). PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. *Bioinformatics*, *21*(16), 3445-3447.

Wing, J. K., Babor, T., Brugha, T., Burke, J., Cooper, J. E., Giel, R. et al. (1990). SCAN: Schedules fonr Clinical Assessment in Neuropsychiatry. *Archives of general psychiatry*, *47*(6), 589.

Winokur, G., Coryell, W., Keller, M., Endicott, J., & Leon, A. (1995). A Family Study of Manic-Depressive (Bipolar I) Disease: Is It a Distinct Illness Separable From Primary Unipolar Depression? *Archives of general psychiatry*, *52*(5), 367-373.

Whittemore, A. S., & Halpern, J. (1994). A class of tests for linkage using affected pedigree members. *Biometrics,* 118-127.

Whittemore, A. S. (1996). Genome scanning for linkage: an overview. *American journal of human genetics*, *59*(3), 704-716.

World Health Organization, W. H. (1993). *The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research.* World Health Organization.

Wray, N. R., Goddard, M. E., Larizza, L., Roversi, G., Volpi, L., Boles, R. et al. (2010). Multi-locus models of genetic risk of disease. *Genome Medicine*, *2*(10).

Ylisaukko-oja, T., Nieminen-von Wendt, T., Kempas, E., Sarenius, S., Varilo, T., Von Wendt, L. et al. (2004). Genome-wide scan for loci of Asperger syndrome. *Molecular psychiatry*, *9*(2), 161-168.

Zandi, P. P., Willour, V. L., Huo, Y., Chellis, J., Potash, J. B., MacKinnon, D. F. et al. (2003). Genome scan of a second wave of NIMH genetics initiative bipolar pedigrees: chromosomes 2, 11, 13, 14, and X. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *119*(1), 69-76.

Zhang, D., Cheng, L., Qian, Y., Alliey-Rodriguez, N., Kelsoe, J. R., Greenwood, T. et al. (2008). Singleton deletions throughout the genome increase risk of bipolar disorder. *Molecular psychiatry*, *14*(4), 376-380.

Zheng, G., Freidlin, B., & Gastwirth, J. L. (2006). Robust genomic control for association studies. *The American Journal of Human Genetics*, *78*(2), 350-356.

Zubenko, G. S., Maher, B., Hughes III, H. B., Zubenko, W. N., Stiffler, J. S., Kaplan, B. B. et al. (2003). Genome‐wide linkage survey for genetic loci that influence the development of depressive disorders in families with recurrent, early‐onset, major depression. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *123*(1), 1-18.